Constraint Deviation Engine

Drift Detection and Governance in Stratified Agent Architectures

Stephen A. Putman
putmanmodel@pm.me

Version 0.2 — February 2026

Abstract

Stratified agent architectures require mechanisms that preserve interactional coherence under shifting language, tone, and symbolic context. The Constraint Deviation Engine (CDE) is a drift detection and governance subsystem that detects deviation relative to stratified baselines and detects reference shift and escalation markers as constraint breakage across layered representations, rather than as a single sentiment score. Deviation is computed as a multi-layer vector over lexical, pragmatic, semantic, and affective signals, evaluated against scope-specific baselines (global, per-agent, per-task, and per-scene). These signals are restricted to constraint-accessible observables, preserving non-omniscient operation and preventing inference from ungrounded projection. A gating and hysteresis layer separates transient variance from structurally relevant turning points, while a rationale layer attaches evidence spans and contributing features to support inspection, debugging, and audit. This positions CDE to support early warning of amplification-dominant conditions before reference collapse propagates through downstream planning or multi-agent coordination layers, under explicitly specified precursor criteria.

CDE is designed to interoperate with tiered memory systems by mapping deviation events into short-term volatility, longer-horizon consolidation candidates, and rare prior-level adjustments. Baseline adaptation is regulated to reduce uncontrolled assimilation and preserve continuity under adversarial or unstable inputs. We present CDE as a modular pipeline—normalization, signal extraction, baseline retrieval, deviation computation, gated classification, rationale generation, routing, and guarded update—and discuss deployment patterns for conversational agents, multi-agent simulations, and human-in-the-loop systems. The result is a disciplined account of drift as enforceable, reviewable constraint deviation within time-extended agent operation.

1. Introduction

Artificial agents increasingly operate in settings where language is not merely descriptive, but interactional: utterances modulate reference, pressure, coordination, and downstream action. In such settings, failure is often preceded by a detectable transition: a shift from distributed reference and reciprocal coordination into concentrated reference, escalation, or symbolic conflict. CDE is introduced as a governance closure mechanism for stratified agent architectures: it detects and structures deviation as an explicit, auditable event that can be routed into control actions and memory governance.

Position in trilogy. Constraint-grounded inference restricts what an agent may justifiably infer from accessible structure, distinguishing traceable inference from projection beyond invariant boundary effects. Memory stratification defines persistence as time-scaled representational continuity, regulated through tier-specific write resistance and gated assimilation. Multi-agent field dynamics formalize collective instability as reference shift and cascade under amplification-dominant conditions relative to dissipation capacity. CDE closes the arc by detecting precursor deviation conditions and routing them into governed responses before collapse propagates through planning, coordination, or consolidation pathways.

CDE is architectural rather than model-specific. It does not presume a particular language model, classifier, embedding method, or emotional schema. Instead, it defines (i) a constrained observables boundary, (ii) a stratified baseline structure, (iii) an event formation mechanism, and (iv) governance interfaces for memory and coordination layers.

2. Terminology and Operational Definitions

All terms are used in an operational and architectural sense.

Constraint-Accessible Observables

Measurable signals available to the agent through its interaction surface, including text, metadata, and bounded derived features computed from them. Derived features are permitted only if computed strictly from the provided interaction payload and its explicit metadata. Hidden model internals and external world lookups are disallowed unless surfaced as explicit constraint-accessible observables through the same interface.

State (Abstraction Space)

Let $S(t)$ denote a generic state vector over constraint-accessible abstractions derived from interaction. CDE operates on $S(t)$ and derived deviation signals; it does not require direct access to raw observables, and it does not depend on any specific representational schema.

Baseline

A baseline is a scope-bound reference model defined over constraint-accessible abstractions derived from interaction. It preserves invariants common to the relevant constraint surfaces and serves as the reference against which deviation is computed. Baseline construction is restricted to representations derivable from interaction, ensuring that deviation is evaluated in abstraction space rather than at the level of raw observables. In this specification, “baseline” refers only to the scope-bound reference model used for deviation computation. Minimum-activity floors or stabilization terms used in other architectures are out of scope and are not treated as baselines here.

Deviation Vector

A layered set of distance measures between current abstraction-state signals and baselines, computed across multiple signal layers (lexical, pragmatic, semantic, affective). The vector form is retained for routing and rationale.

Reference Shift

A measurable transition in interactional orientation, expressed through observable markers indicating movement away from shared constraint alignment and reciprocal exchange.

Escalation Marker

Observable indicators of intensified interactional pressure (e.g., emphasis amplification, hostile lexical transitions, mock-echo patterns), treated as constraint deviation signals rather than as psychological claims.

Event Formation

The transformation of a deviation vector into a discrete deviation event through gating, persistence, and hysteresis, separating transient variance from structurally relevant turning points.

Rationale Artifact

A structured explanation attached to a deviation event, including evidence spans, baseline comparison, contributing features, and extractor/version provenance sufficient for replay.

Guarded Update

Regulated baseline modification that preserves continuity by constraining when and how baselines assimilate new observations.

3. Constraint-Accessible Observables and Signal Extraction

CDE begins by defining what may be measured. Observables are restricted to interaction-accessible signals and bounded derived features computed from them. This preserves non-omniscient operation and prevents drift detection from relying on hidden-state claims or post hoc narrative attributions.

Each turn is represented as a minimal packet: (i) text (or interaction payload), (ii) timestamp, (iii) speaker/agent identifier, (iv) channel identifier, and (v) optional bounded context references (task id, scene id, policy state). No external world-state is assumed unless it is explicitly provided as a constraint-accessible observable through the same interface.

CDE supports multiple signal layers, computed in parallel and treated as distinct observables over abstraction state:

- Lexical layer: token-level markers, emphasis punctuation patterns, casing intensity, repeated symbols.

- Pragmatic layer: request–demand transitions, hedging contraction, stance rigidity markers.
- Semantic layer: bounded similarity change relative to prior state baselines or topic anchors.
- Affective layer: bounded tone-vector estimation derived from interaction-accessible features.

CDE does not require that each layer be present. The engine is defined by the computation of a deviation vector over whatever layers are available, with explicit confidence bounds.

Each signal layer produces evidence spans: localized fragments supporting the computed signal. Evidence spans are first-class outputs for inspection and audit and must be reproducible given the same inputs and extractor versions. A minimal evidence span schema includes: span_id, turn_id, layer, char_range (and/or token_range), score, confidence, attribution_method_id, extractor_version, and seed (only if stochastic attribution is used). The determinism contract is: identical input payload + identical extractor_version (+ identical seed when applicable) yields identical span outputs.

4. Stratified Baselines and Deviation Computation

CDE computes deviation relative to stratified baselines. A baseline is a scope-bound reference model defined over constraint-accessible abstractions derived from interaction; it preserves invariants common to the relevant constraint surfaces and provides the reference frame in which deviation is evaluated. Baselines are maintained by scope to avoid collapsing heterogeneous context into a single global distribution, and deviation is computed in abstraction space rather than at the level of raw observables.

The minimal baseline set includes: global baseline (system-wide operating distribution), per-agent baseline (speaker-specific operating distribution), per-task baseline (task or domain distribution), and per-scene baseline (interactional field / simulation context distribution). Scope selection is not exclusive. CDE may compute deviation against multiple baselines simultaneously and report which scope(s) produce meaningful deviation.

Each baseline $B^{(k)}$ carries a manifest sufficient for deterministic retrieval and audit. The manifest minimally includes: scope key k, baseline family identifier, parameter/version identifier, manifest_version, and baseline_hash. Required lifecycle operations are: (i) retrieve a baseline deterministically for a given (scope key, identity key, time), (ii) freeze/lock a baseline under governance modes, (iii) apply guarded updates when eligible, and (iv) record manifest changes in the audit log. Baseline selection for a given $S(t)$ must be determined solely by constraint-accessible metadata (e.g., agent id, task id, scene id) and must be reproducible under replay.

To avoid circular dependency at initialization, CDE supports a warm-up mode in which baselines and weights are read-only and only monitoring and logging occur. Warm-up is active for a bounded horizon H_0 declared in the manifest; after H_0, guarded updates may be enabled under the standard gating and confidence regimes. Any transition into or out of warm-up mode is recorded in the audit log.

The baseline representation is implementation-dependent, but must belong to an admissible family with a well-defined distance contract. Examples include: robust summary statistics over a feature space (e.g., location/scale), bounded exemplar sets with a similarity kernel, or centroid-plus-dispersion summaries for embedding-like layers. The paper constrains the admissible design space without prescribing a single family.

For each scope k, deviation is defined as the layerwise distance between the current abstraction-state and the scope baseline. Let $S(t)$ denote the abstraction-state at time t. Let $B^{(k)}$ denote a baseline model for scope k Let $\mathscr{L}$ be the set of signal layers (lexical, pragmatic, semantic, affective). For each layer $\ell \in \mathscr{L}$, define a distance function $d_\ell(\cdot, \cdot)$. The deviation vector is:

$$\Delta_t^{(k)} = \left[d_\ell(S_{t,\ell}, B_\ell^{(k)})\right]_{\ell \in \mathscr{L}}$$

For governance gating, the deviation vector is summarized into an aggregate severity under scope-declared weights. Canonically,

$$A_t^{(k)} = \|\Delta_t^{(k)}\|_{\mathbf{w}} = \sqrt{\sum_{\ell \in \mathscr{L}} \mathbf{w}_\ell^{(k)} (\Delta_{t,\ell}^{(k)})^2}, \quad \mathbf{w}_\ell^{(k)} \geq 0, \ \sum_{\ell} \mathbf{w}_\ell^{(k)} = 1$$

Alternate aggregation norms are admissible if declared in the manifest.

To ensure comparability across heterogeneous layers, each d_ℓ must produce a calibrated scalar with documented semantics on a shared normalized range (by default [0,1]). If a native distance is unbounded or not directly comparable, it must be mapped through a layer-specific normalization transform N_ℓ prior to aggregation. Thresholds $\theta^{(k)}$, persistence rules, and routing conditions are defined on the normalized aggregate $A_t^{(k)}$. Weight vectors $\mathbf{w}^{(k)}$ are scope-specific governance parameters stored in the baseline manifest; online weight learning is non-normative and, if used, must occur via an explicit manifest version change and be logged.

Confidence is aggregated analogously to severity to support deterministic routing under uncertainty. Let $c_{t,\ell} \in [0,1]$ denote per-layer confidence values. Canonically, aggregate confidence for scope k is:

$$C_t^{(k)} = \sum_{\ell \in \mathscr{L}} \mathbf{w}_\ell^{(k)} c_{t,\ell}$$

Confidence is discretized into Low/Medium/High regimes using two thresholds , with $c_{\text{low}}^{(k)} < c_{\text{high}}^{(k)}$, declared per scope in the manifest and logged for replay.

When one or more signal layers are absent at time t, those layers are excluded from aggregation and the remaining weights $\mathbf{w}^{(k)}$ are renormalized proportionally over the present layers for both $A_t^{(k)}$ and $C_t^{(k)}$. The missing-layer policy is declared per scope in the manifest and logged for replay.

CDE reports scope-triggered deviation rather than forcing a single interpretation. Scope arbitration returns a ranked list of triggered scopes with severities and rationale. A deterministic comparator is required; canonically, scopes are sorted by (A desc, C desc, $p(k)$ asc) where $p(\text{per-scene}) = 1, p(\text{per-task}) = 2, p(\text{per-agent}) = 3, p(\text{global}) = 4$. Any deviation from the canonical comparator must be recorded in the manifest.

In some architectures, it is useful to represent higher-level operating targets as constraints in a bounded goal-space. One canonical example is a bounded simplex (triangle/tetrahedron/etc.) used as a goal manifold. In such a framing, events (local tiles) and sustained trends (modes) are projected into goal space, and deviation is evaluated as constraint violation (outside-simplex) or divergence from a target mixture. Compact metrics include: (i) goal_distance (point-to-simplex or point-to-constraint-set distance), (ii) goal_divergence (divergence between current mode mixture and target mixture), and (iii) trajectory_angle (angle between current drift vector and nearest admissible or goal-aligned direction). This framing is illustrative and is not a dependency of CDE.

CDE does not claim intent, belief, or hidden state. It reports deviation in constraint-accessible abstractions and flags structural transitions consistent with reference shift and escalation markers.

5. Event Formation: Gating, Persistence, Hysteresis, and Turning-Point Detection

Raw deviation does not constitute a governance event. Event formation distinguishes transient variance from structurally relevant turning points.

For each scope k, define an entry threshold $\theta^{(k)}_{\text{enter}}$. Let $A^{(k)}_t$ be the normalized aggregate severity. Event eligibility is defined under a required persistence rule: deviation is accumulated using a documented accumulator and must exceed $\theta^{(k)}_{\text{enter}}$ under the applicable confidence regime. Canonically, an exponential moving average (EMA) accumulator is used:

$$\widehat{A}^{(k)}_t = (1 - \beta^{(k)})A^{(k)}_t + \beta^{(k)}\widehat{A}^{(k)}_{t-1}, \quad \beta^{(k)} \in (0,1)$$

Larger $\beta^{(k)}$ increases persistence (slower forgetting). The parameter $\beta^{(k)}$ (and optionally an effective window $W^{(k)}$) is declared per scope in the manifest and logged for replay. Deviations from the canonical persistence rule are permitted but must be recorded in the manifest and audit log.

To prevent flip-flop governance, CDE uses distinct entry and exit thresholds: enter event if $\widehat{A}_t^{(k)} \geq \theta_{\text{enter}}^{(k)}$; remain in event-state until $\widehat{A}_t^{(k)} \leq \theta_{\text{exit}}^{(k)}$. A canonical specification is $\theta_{\text{exit}}^{(k)} = \alpha^{(k)}\theta_{\text{enter}}^{(k)}$ with $\alpha^{(k)} \in (0,1)$ documented per scope in the baseline manifest.

Once gated, CDE assigns a deviation class based on dominant contributing layers and reference shift markers. Classes are operational (e.g., escalation marker dominant, inversion dominant, scope-discrepancy dominant) and are tied to routing contracts, not interpretive narratives.

The following primitives are treated as governance hardening rather than optional tuning: (i) hysteresis for exiting safety modes, (ii) rate limits (viscous resistance) on how rapidly modes and constraints may shift, (iii) an outlier quarantine pool that stores deviation-bearing items while preventing them from driving baseline updates or responses, and (iv) conservative arbitration for high-impact actions (if analyzers disagree or confidence is low, route to the safest path and/or review band).

Quarantined items are excluded from baseline update statistics by default; inclusion is non-normative and, if enabled for a specific baseline family, must be declared in the manifest.

CDE is positioned to support early warning of amplification-dominant conditions when precursor criteria are explicitly specified over normalized deviation trajectories and marker densities. In this paper, precursor detection is treated as an architectural interface: CDE requires that precursor flags be definable as reproducible predicates over (i) sustained normalized deviation and (ii) reference shift and escalation marker measures, and that precursor evaluation be reported under falsifiability tests rather than asserted categorically.

6. Governance Routing in Stratified Architectures

CDE outputs are governance inputs. A deviation event produces a rationale artifact and a routing decision. Routing is deterministic at the interface level, even when upstream signal extractors are probabilistic.

Each event yields: deviation vector and severity $A_t^{(k)}$ (and $\widehat{A}_t^{(k)}$ where applicable), triggered scope(s), event class, rationale artifact (evidence spans, contributing features, baseline deltas, extractor/version provenance), and a recommended governance action set.

The minimal action families are: policy gating (restrict or expand permitted downstream actions under deviation conditions), interaction constraint adjustment (enforce de-escalation or clarification protocols), escalation routing (trigger human-in-the-loop review or supervisory arbitration), coordination dampening (reduce coupling or influence weights in coordination layers under precursor conditions), and quarantine (suspend baseline updates or memory consolidation under unstable periods).

Extractors may output probabilistic signals and confidences; routing remains deterministic by operating on discrete severity and confidence regimes. The canonical routing input is the pair $(\widehat{A}_t^{(k)}, C_t^{(k)})$ for each triggered scope k. Low-confidence regimes route to conservative governance modes (quarantine and/or review band) rather than forcing high-impact actions. Confidence thresholds and band definitions are governance parameters and must be versioned in the manifest.

CDE may implement graded interventions as layered gates that block instability from propagating upward through the architecture: Gate A (Input / Step 1) quarantines out-of-gamut events, reduces influence, and forces mediation templates or constrained response forms; Gate B (Trend / Step 2) treats sustained divergence as a mode-lock condition, freezes goal updates, and limits coordination influence until dissipation conditions return; Gate C (Archetype / Step 3) treats goal definitions as write-protected (signed and versioned) and prohibits mutation by lower gates, requiring explicit governance pathways for changes. This gate-stack is described as a governance pattern rather than a mandatory subsystem.

CDE produces a minimal audit log sufficient for review and replay. Required fields include: event id, timestamp, scope(s), class, severity regime, confidence regime, evidence span references, baseline_hash, manifest_version, extractor_version(s), attribution_method_id(s), and action outputs. If stochastic attribution is used, the relevant seed (or a pointer to stored evidence-span records containing the seed) must be included to guarantee replayability. If the optional goal-manifold framing is used, include: goal_distance, goal_divergence, trajectory_angle. For layered gates, include: lock_state (None | ModeLock | ArchetypeFreeze), quarantined_items_count (optionally last N ids), goal_manifest_hash, manifest_version, signer.

7. Memory-Governed Drift: Tier Mapping and Commit Candidates

CDE is designed to interoperate with stratified memory architectures in which tiers differ in persistence, write resistance, and assimilation gating.

Deviation events are mapped to memory-tier relevance: surface volatility candidates (deviation events that alter near-term response constraints but do not justify consolidation), intermediate consolidation candidates (repeated or sustained deviation patterns that represent compressed regularity relevant to future constraint handling), and prior-level adjustment candidates (rare, low-frequency patterns that justify modifying deep priors). This mapping is governed. CDE proposes commit candidates; the memory system admits or rejects them according to tier permeability and write resistance.

Stratified memory produces residual activation when activated structure is not yet integrated into longer-horizon tiers. CDE provides a structured mechanism for distinguishing transient activation spikes from repeatable deviation patterns that merit consolidation candidacy, reducing uncontrolled accumulation of unresolved pressure.

8. Guarded Update and Baseline Integrity

Reference baselines may be updated to maintain relevance, but uncontrolled updates collapse scope distinctions and degrade deviation meaning. CDE therefore regulates baseline adaptation via guarded update rules that preserve continuity under unstable sequences.

Let $B^{(k)}$ be a baseline for scope k. Let $u_t^{(k)} \in [0,1]$ be a scope-local update gate determined by event state, severity regime, precursor flags, and audit policy. Baseline adaptation is governed by a gated update rule that is suppressed under instability:

$$B_{t+1}^{(k)} = (1 - \eta u_t^{(k)})\, B_t^{(k)} + \eta u_t^{(k)}\, \phi(S(t))$$

where η is a learning rate and $\phi(S(t))$ is a bounded update statistic derived from constraint-accessible abstraction state, consistent with the baseline family in the manifest.

As a baseline-family constraint, this update rule applies only to baseline families that admit interpolation within a shared parameter space (e.g., centroid or summary-statistic families). For non-interpolable families (e.g., exemplar/kernel baselines), guarded update uses a family-specific bounded operator (e.g., insertion/retirement under a declared reservoir policy) recorded in the manifest.

Minimal $u_t^{(k)}$ decision contract. The mapping from event regime and governance mode to $u_t^{(k)}$ must be explicit and logged. One canonical banded mapping is: high severity or active precursor flags $\Rightarrow u_t^{(k)} = 0$ (freeze); medium severity $\Rightarrow u_t^{(k)} = u_{\text{mid}}$ (partial update); low severity / stable $\Rightarrow u_t^{(k)} = 1$ (update). Canonically $u_{\text{mid}} = 0.5$ unless overridden in the manifest.

Multi-scope update gating. When multiple scopes are evaluated concurrently, baseline updates remain scope-stratified: narrow-scope instability freezes learning in that scope and blocks upward assimilation into broader scopes. For any shared update decision that must be made across scopes (e.g., shared parameters or coupled baselines), the canonical aggregation is conservative: $u_t = \min_k u_t^{(k)}$ over triggered scopes.

Persistent freeze promotion (governance closure). Freeze is treated as immediate integrity protection. Repeated freeze triggers within a bounded horizon H for a given scope (not necessarily contiguous) are promoted to a governed mode transition (ModeLock) that blocks upward assimilation and requires explicit baseline manifest versioning or escalation routing (e.g., Gate B), rather than indefinite freeze without resolution. The horizon H and promotion threshold are declared in the manifest.

Guarded update is treated as an architectural requirement analogous to regulated permeability in memory stratification: excessive assimilation collapses structure; excessive resistance produces update lag and misalignment accumulation.

9. Architectural Implications

CDE imposes design constraints for stratified agent stacks.

1. Constraint-accessible deviation only. Drift signals must be derived from interaction-derived abstractions; non-omniscient operation is preserved by construction.
2. Scope stratification is non-optional. Global-only baselines collapse context and reduce governance fidelity; per-agent and per-scene baselines are required for time-extended coherence.
3. Event formation requires hysteresis and persistence. Governance must operate on turning points, not raw variance; hysteresis and persistence rules are required to prevent oscillatory control loops.
4. Rationale artifacts are part of the contract. Without evidence spans and provenance, deviation events cannot be audited or debugged at architectural scale.
5. Precursor routing must reach coordination layers. If deviation detection is downstream of planning/coordination, the system detects collapse after propagation rather than before it.

Non-normative note on noticeability. Architectures may optionally expose observability hooks (e.g., a compact deviation indicator with short tail, quarantine markers, or sustained-drift overlays) without making UI semantics normative to CDE.

10. Falsifiability and Concrete Test Classes

CDE may be empirically challenged through constrained tests with clear pass/fail criteria.

1. Inversion tests. Sarcasm, negation, quoted speech, and reversal markers should produce layer-appropriate deviation and evidence spans without spurious escalation classification, with bounded false-alarm rate under the declared confidence regimes.
2. Escalation ramp tests. Controlled sequences from neutral → pressure → hostile should produce gated turning-point detection before peak escalation, with detection lead-time reported and hysteresis limiting oscillation count under repeated perturbation.
3. Scope arbitration tests. The same utterance in different scenes/tasks should yield different deviation outcomes when baselines differ, demonstrating scope sensitivity with an explicit scope-consistency rate under replay.
4. Baseline integrity tests. Under unstable sequences, guarded update should prevent rapid assimilation and preserve deviation detectability when stability returns, with recovery time reported after freeze periods.
5. Precursor tests in multi-agent coordination. Under simulated amplification-dominant conditions, declared precursor predicates should be detectable before reference collapse propagates through coordination layers, with lead-time distribution and false-precursor rate reported.

11. Conclusion

Constraint-grounded inference defines what an agent may justifiably infer from accessible structure and prohibits extension beyond invariant support as projection. Memory stratification defines how an agent persists across time through regulated coupling among tiers of differing write resistance. Collective field dynamics define how multi-agent systems destabilize through reference shift and cascade under amplification-dominant conditions relative to dissipation capacity. CDE closes the architectural arc by detecting deviation as constraint breakage in stratified, interaction-derived abstractions, forming governed events with auditable rationales, and routing these events into policy, memory, and coordination controls before collapse propagates. The result is an enforceable account of drift as reviewable constraint deviation within time-extended, stratified agent operation.