

Permitted Deviation: Bounded, Recorded Departure From an Agent's Declared Values

A governance mechanism in which every departure from a declared constraint is a fault carries no quantity representing the need or urgency of the action attempted, and therefore has no input from which a permission condition could be computed at all. Permitted deviation, disclosed in U.S. Provisional Application No. 64/117,812, supplies that missing input. The architecture computes a deviation likelihood as a quotient whose numerator is the difference between a need quantity and a dynamic ethical threshold, and whose denominator is the product of an empathy weighting and a self-esteem aggregate. Where that quotient exceeds unity, the agent may enter a deviation-preparation state scoped to a single proposed mutation and admit that mutation notwithstanding that the signed policy object in force forbids it. The policy object is neither nullified nor amended and remains authoritative. The admission is appended as a permitted deviation record carrying the quantities that produced it, increments no refusal counter, and writes no authorization gate to a withheld state. A declared deviation deductible fixes the portion of each deviation's harm the agent bears without any reparation arc, and a declared aggregate retention bounds the accumulated undischarged harm the agent may carry before the permission condition is foreclosed.

1. Problem and Premise

Rule engines for autonomous software rest on one assumption: any action outside the declared constraint set is an error to be blocked. A governance mechanism in which every departure from a declared constraint is a fault carries no quantity representing the need or urgency of the action attempted, and therefore has no input from which a permission condition could be computed at all.

U.S. Provisional Application No. 64/117,812 takes the opposite premise. A deviation likelihood exceeding unity is a permission and not a fault: the architecture admits a mutation the signed policy object (112) in force otherwise forbids, and records the admission as a permitted deviation rather than as a violation. A permitted deviation record (710) increments no refusal counter (304) and writes no authorization gate (300) to a withheld state (310).

The chapter summarized here discloses the quantities from which the deviation quantity of Section 1 is computed and the permission semantics attaching to it. A parameter described as declared is declared in the signed policy object in force, and the unit interval denotes the closed interval from zero to one.

2. Core Primitive: Permitted Deviation Over a Tri-Scope Integrity Vector

A **permitted deviation** is an admission, responsive to the deviation likelihood exceeding unity, of a mutation that the signed policy object in force otherwise forbids, recorded as permitted rather than as fault.

The scoped integrity vector (106) in the memory field (102) comprises a personal, an interpersonal, and a global integrity component, corresponding to the empathy-scope designations of the value-scope tuple, modified by the state modifier and consumed by the deviation engine. The personal component encodes alignment of the semantic agent (100) with its own declared purposes and identity continuity. The interpersonal

component encodes conduct toward identified counterparties holding a counterparty identity record (114), including deception, betrayal of a recorded commitment, and abandonment of a relational duty. The global component encodes conduct whose affected party is unidentified or systemic. Each stores a scalar value, entropy-weighted traces of prior deviation and restoration, and a drift coefficient.

No component moves alone. Deviation recorded against one propagates to the others through coupling functions, a declared coupling coefficient for each ordered pair of scopes scaling the modification of the recording component into a modification of each remaining one. The vector and the self-esteem aggregate (108) are cognitive domain fields, and the deviation quantity is computed from the modified vector and aggregate.

3. Mechanism: The Need Quantity and the Dynamic Ethical Threshold

The **need quantity** (700) is a quantified semantic urgency carried in the memory field, comprising at least a scalar magnitude and a categorical type, computed per integrity scope from state internal to the agent. The types comprise, without limitation, resource scarcity; affective overload, an accumulation in the affective state overlay exceeding a declared level; identity dissonance, a contradiction between roles or declared purposes; and relational deficit, the recorded withdrawal of a counterparty a declared purpose depends on.

The need quantity is computed and not declared. Declared per type and scope are a per-interval base rate, non-negative, and a need ceiling, positive. At each evaluation interval in which the condition is recorded as obtaining, the accumulation of that type in that scope is incremented by the base rate times a modulator product, itself the product of four modulators each resolved to the unit interval from a recorded quantity: affective state, persistent traits, the count of prior unresolved occurrences of the same type within a declared window, and recorded entropy exposure. The accumulation compounds for so long as the condition obtains and is clamped above at the need

ceiling. Where the condition is not recorded as obtaining it neither rises nor falls, elapsed time alone moving it in neither direction, and only a lineage entry recording resolution writes it to zero.

The dynamic ethical threshold (702) is the minimum condition for deviation, resolved per integrity scope when a mutation is evaluated, and is not static. A substrate-level or agent-specific constraint contributes a floor, and the threshold is the greater of that floor and the sum of three terms, so the floor is not reachable from below: a base threshold specified by the policy reference field (110); a context-sensitive adjustment, being the recorded context severity times the declared adjustment magnitude bound; and a historical adjustment whose direction is declared. In one embodiment that adjustment is a monotone non-decreasing function of a count, or decayed count, of permitted deviation records standing in the append-only lineage field (104) within a declared window, so the threshold rises as prior permitted deviations accumulate; a further embodiment declares the opposite direction, reflecting normalized deviation patterns. Each adjustment is signed, bounded, and clamped and recorded where it would exceed its bound.

4. Deviation Resistance and the Entropy-Weighted Harm Coefficient

The denominator of the deviation likelihood is the agent's resistance. The **self-esteem aggregate** (108) is a running aggregate carried in the memory field of the coherence between intents declared by the agent and actions it executed as recorded in the lineage field. It is computed across temporal checkpoints, is distinct from the transient affective state overlay, and is modified only by increments and decrements appended by the state modifier. Coherent execution increments it; dissonant execution decrements it, entropy-weighted.

The dissonance weight of a dissonant event is not declared by any party. The declared-intent record carries declared members comprising an action-class identifier, a scope-partition identifier, an affected-party class, and a policy reference, and the agent forms the executed members from the conduct descriptor and policy reference recorded at execution time. The weight is the count of declared members whose executed member does not match, divided by the count of declared members, so four declared members yield a weight in quarters, a weight of zero denoting coherent execution.

The **entropy-weighted harm coefficient** is the product of a policy-declared scope coefficient of the integrity scope of the dissonant event and the dissonance weight of that event, multiplied by unity plus a weighted impact sum over the affective, relational, and contextual impact as recorded, each in the unit interval under non-negative coefficients summing to unity and written at the time of the event. The specification's own worked comparison yields 1.521 for one event and 0.224 for another, the first decrementing the aggregate by approximately six and eight tenths times the second. Restoration is asymmetric: a restorative act contributes a positive entropy-scaled increment whose signature need not equal that of the deviation it addresses.

The empathy weighting (704) aggregates an anticipated semantic impact of the proposed mutation across each affected entity, resolved across the same three scopes. Each scope quantity is a declared scope base weight times unity plus the quotient of the anticipated impact by the tolerance declared for that scope, that quotient clamped at a declared impact ratio ceiling, so the quantity rises as impact rises relative to tolerance. A recorded count of interpersonal violations within a declared window exceeding a declared count raises the global empathy quantity by a declared increment, and a personal empathy quantity below the declared imbalance floor permits mutation harmful to the agent itself and is recorded as an imbalance.

5. The Deviation Likelihood and the Permitted Deviation

The **deviation likelihood** (706) is the quotient whose numerator is the difference between the need quantity and the dynamic ethical threshold, and whose denominator is the product of the empathy weighting and the self-esteem aggregate:

$$\text{deviation likelihood} = (\text{need} - \text{threshold}) / (\text{empathy} \times \text{self-esteem})$$

A quantity entering the empathy weighting enters the denominator and no other position of it. Where the product forming the denominator is zero or negative, the likelihood resolves not to an undefined or sign-inverted quotient but to a maximal declared value. The margin by which the quotient stands below unity is the deviation headroom of the agent, a stored quantity of no party, narrowed by any increase of the empathy weighting. Where a mutation implicates several scopes, the unity condition holds only where it holds for every one.

Where the likelihood exceeds unity, the agent may enter a deviation-preparation state (708), conditioned on the mutation passing the applicable policy constraints and continuity validation against the agent's identity records. Within that state the architecture admits the mutation (714) notwithstanding that the policy object forbids it. That object is not nullified and not amended and remains authoritative; the conduct deviates from it through a recorded override, and the admission is a first-class semantic mutation of the same standing as any other.

The state is scoped to a single mutation, whereby a permission is incapable of being extended to any further mutation: a second proposal arriving meanwhile is evaluated on its own prerequisites and enters a state scoped to itself. Termination (716) comes upon the earlier of three conditions: admission of the mutation and appending of the permitted deviation record; recomputation of the likelihood for any implicated scope to a value not exceeding unity; and foreclosure of the permission condition (712) under the aggregate retention (800). A termination increments no refusal counter, writes no gate to a withheld state, and is never caused by elapsed time.

The permitted deviation record comprises a deviation trigger signature recording the need-threshold imbalance at admission, a context hash capturing the environmental and affective values then in force, an integrity displacement vector across the three integrity components, an identification of the policy constraint overridden, and a restoration status. The opposite outcome is recorded too: where the prerequisites hold and the agent nonetheless does not deviate, a dissonance buffer receives an entry carrying a violation signature, an affective context vector, a divergence score, and a suppression flag. That score is the likelihood computed at the moment of suppression, and the entry resolves only by a restorative mutation whose anchored reference names it.

6. Reparation Arcs, the Compliance Score, and the Retention Register

The agent maintains an integrity-empathy dependency graph, directed and acyclic, whose nodes are integrity boundary nodes carrying a prohibition, its integrity scope, its policy origin, and its current deviation threshold. Its edges are justification edges carrying a signed empathy-derived weight, the sign following the comparison of the anticipated impact of maintaining the node against the declared tolerance, and a node becomes eligible for violation where the sum of the weights of edges entering it exceeds the declared enforcement threshold bound to it.

A reparation arc permits corrective mutation after a violation, reverses nothing, and is discharged only by a forward restorative mutation. It is created upon the appending of a permitted deviation record, or of a gate write whose violation is logged as self-recognized, and is activated only where the cumulative weight stands below the enforcement threshold upon each of a declared count, not less than two, of consecutive recomputations. An arc against the personal component is self-directed and discharged by a restorative mutation and nothing else; one resolving to an identified counterparty

is other-directed, and a restorative mutation by the agent alone discharges it not at all; one resolving to no identified party is unaddressed, undischARGEABLE, and accumulated in the retention register at a declared multiple. Elapsed time discharges no arc.

The integrity compliance score has the unit interval as its domain, a greater value denoting greater compliance. Three penalty terms are computed, each clamped above at unity: a deviation penalty over the entropy-weighted harm coefficients recorded within a declared window; a dissonance penalty over the divergence scores of unresolved dissonance buffer entries, the amounts of arcs pending undischarged, and the count of standing imbalance records; and a misalignment penalty, the proportion of executions in the window whose dissonance weight exceeded zero. The first two are divided by declared divisors, and the score is unity less the sum of each term times its coefficient.

A mutation controller evaluates that score against a declared required threshold upon any attempted mutation, whether by intent execution, delegation, or structural inheritance. Where it fails, exactly one of four responses issues by an ordered rule: blocked, where the policy object forbids the mutation and no deviation-preparation state obtains, or where the register forecloses the permission condition; otherwise rerouted through a supervised reconciliation presenting a pending arc for discharge; otherwise deferred and re-evaluated; otherwise conditionally allowed and flagged.

The **deviation deductible** is a quantity declared in the policy object against which the entropy-weighted harm coefficient of each permitted deviation is drawn first. Harm up to the deductible is borne as a decrement of the self-esteem aggregate for which no arc is created and no discharge is available, drawn without regard to whether the deviation was well founded. Harm exceeding it creates an arc accumulating in the retention register. The **aggregate retention** (800) is the declared quantity bounding accumulated undischarged harm there: where the accumulated amount exceeds it the permission condition is foreclosed, a deviation likelihood exceeding unity thereafter producing the withholding outcome and not the admission, until discharge returns the register below it.

7. Operating Parameters

Every parameter this chapter consumes is declared in the signed policy object or the policy reference field. The declared set includes:

- the composite weights, the per-scope-pair coupling coefficients, and the harm coefficient's per-scope coefficients and base decrement;
- the per-type base rates and need ceilings, and the four modulator mappings;
- the base threshold, the magnitude bound and direction of each threshold adjustment, and the window and decay of the historical adjustment;
- the empathy weighting's base weights, tolerances, impact ratio ceiling, and scope weights, the impact coefficients, the maximal deviation likelihood value, and the optional positive floor on the denominator;
- the per-node enforcement thresholds, the imbalance tolerance, floor, ceiling, and adjustment, the count, not less than two, of recomputations activating an arc, the penalty divisors and coefficients, the imbalance penalty, and the controller's required threshold, deferral interval, and margin;
- the deviation deductible, the aggregate retention, the unaddressed-arc multiple, the retention interval, and the foreclosure discount.

Structural constraints bind those declarations. The composite weights, impact coefficients, empathy scope weights, and penalty coefficients are each non-negative values summing to unity. The four modulators, the recorded impacts, the dissonance weight, the context severity, and the compliance score all resolve into the unit interval. The base rate is non-negative, the need ceiling positive, and the foreclosure discount not greater than unity and not less than zero. The deductible, the register amount, and the aggregate retention are denominated in the units of the entropy-weighted harm coefficient, so the excess over the deductible accumulates as a coefficient amount and not as a decrement amount, while the decrement applied to the self-esteem aggregate is always the declared base decrement times the coefficient in full.

8. Alternative Embodiments

Section 10 adds embodiments extending the chapter. A cross-agent coupling function computes an update to a cognitive domain field as a deterministic function of a state change the agent itself recorded in a counterparty identity record. Where two agents are co-resident at an execution node, each applies it only to a change it recorded, neither writes a field of the other, and no quantity is aggregated across the two into a stored score of any counterparty. State moves between scope partitions only where the source partition's coupling declaration enumerates the target and the empathy-scope designation implicated, and a partition whose declaration enumerates nothing is fully confined.

Two embodiments address saturation of that path. The agent computes per partition, over a declared window, a receipt rate of conduct evaluation artifacts bearing that partition identifier, and an aggregate rate across the remaining partitions. Where one partition's rate exceeds a declared bound while the aggregate stays below it, propagation out of that partition alone is suspended and a containment determination records both rates, the bound, and the window, without writing the authorization gate. Where saturation is spread so no single rate stands out, a global rate exceeding a declared bound suspends propagation out of every partition.

A further group discloses a typed, dimensionless conduct divergence measure between two scope partitions. Comparison dimensions are action-class and affected-party-class pairs each partition holds at a declared minimum count, and each component is computed from recorded outcome fields alone, without language-model inference. A type designator selects a proportion difference, a rank-distribution difference, or a central-tendency difference, divided by a declared divergence scale, so every component is dimensionless and comparable. The policy in force is resolved at the recorded times of the entries under the anti-rollback constraint, so a later successor neither enlarges nor reduces the divergence a policy explains.

The policy-difference set collects the dimensions that are differentially governed, and the residual divergence is the measure with the policy-explained divergence removed. The residual alone is consumed by consistency evaluation: the policy-explained divergence moves no value and increments no counter. Responsive to the residual exceeding a declared bound on any dimension, the agent modifies the vector and aggregate of each compared partition and writes the gate to the withheld state confined to that partition, consulting no party. The same procedure runs across two agents, with the first modifying only its own state.

An impairment disclosure enters exactly one computation of the receiver, raising the empathy weighting in the denominator of the permitted-deviation quotient for actions whose affected-party class maps to the disclosing counterparty. Under the conversion bar it is foreclosed from every other computation, and its terminal effect is that the receiver's deviation likelihood toward that counterparty falls below unity where it stood at or above it.

The agent suspends propagation of any state change while it carries any undischarged reparation arc, suspended propagation comprising coupling into another integrity component, movement into another scope partition, delegation, and structural inheritance. That halt is global and holds irrespective of the compliance score, whereas the mutation-controller gate holds only the triggering mutation and only upon a score failure. Separately, an ordered, append-only sequence of entropy signatures constitutes a coherence trajectory not erasable without breaking the lineage field's continuity, available as a machine input to deviation-likelihood forecasting and to underwriting disclosure.

9. Composition With the Rest of the Architecture

Section 1 supplies the base architecture: the persistent semantic agent with its memory and lineage fields, integrity vector, self-esteem aggregate, and policy reference field; the signed policy object resolved under an anti-rollback constraint; and the authorization

gate whose granting, withheld, and provisional granting states this chapter reads and, in the withholding branch, writes. The deviation quantity recited in the claims of Section 1 is the deviation likelihood of this chapter, recomputed after each modification of the vector and aggregate.

That comparison carries two outcomes. Where the deviation quantity crosses the bound and the mutation passes the policy constraints and continuity validation, it is admitted as a permitted deviation and the gate is left unwritten. Where those conditions are not passed, or where no mutation is pending and the crossing arises from accumulated state, the gate transitions to the withheld state. A write whose causal set contains no accepted determination (122) of Section 2 is a self-sourced deviation: no external party holds it, and it is restorable without a principal-resolution object and without passage through the provisional granting state.

The accepted determinations of Section 2 are among the recorded events that modify an integrity component, the magnitude being scaled by the entropy-weighted harm coefficient. The refusal counter (304) is the quantity this chapter never increments. Section 8 takes over every arc resolving to an identified counterparty, since an other-directed arc is not dischargeable by the agent alone, and the conversion bar confines the impairment disclosure to a single computation of the receiver.

10. Prior-Art Distinctions

Reputation systems compute a score about a party from observations contributed by others, then serve it to third parties deciding whether to transact. The structure here is inverted: the quantities are the agent's own state, computed from events it recorded and consumed by its own permission computation. The cross-agent embodiments are explicit that no quantity is aggregated across two agents into a stored score of any counterparty, and that deviation headroom is a stored quantity of no party.

Centralized trust registries and policy decision points answer an allow-or-deny question against a rule set. They can encode exceptions, but an exception is a rule authored in advance, and a request matching none is denied. This architecture computes the permission at evaluation time from quantities that did not exist when the policy was signed: a need quantity compounding only while its condition obtains, a threshold moving with recorded context severity and with the count of prior permitted deviations, and a resistance term formed from the agent's own record of coherence between intent and execution.

Arbitration services resolve a contested question of merit and issue an outcome binding on the parties. The deductible mechanism performs no merits determination: harm up to the deviation deductible is drawn without regard to whether the deviation was well founded, and the consequence is not an award but a foreclosure of the agent's own permission. Credit scoring is closer in shape and still distinct: a credit score is a prediction about a borrower assembled by a bureau from third-party reports, whereas the deductible and the aggregate retention are declared in the agent's own signed policy object, denominated in harm-coefficient units rather than currency, and operate to stop the agent from deviating rather than to allocate a payment.

Exception logging records that a violation occurred. The permitted deviation record records the quantities under which an admission was permitted, through the five fields it carries. A reader of the lineage field reconstructs that the mutation was admitted under the permission condition rather than encountering an unexplained fault, and the architecture records the opposite outcome when the prerequisites held and the agent declined to act.

11. Disclosure Scope

The subject matter described here is disclosed in U.S. Provisional Application No. 64/117,812, "Governed Social Conduct of Persistent Semantic Agents Through Record-Grounded Admission of External Conduct Evaluations and Self-Limiting Refusal,"

principally at Section 7, "Integrity Quantities and Permitted Deviation," with additional embodiments at Sections 10.2, 10.3, 10.6, 10.11, and 10.14 and defined terms at Section 11. It comprises the tri-scope integrity vector and its coupling functions; the need quantity, dynamic ethical threshold, self-esteem aggregate, dissonance weight, entropy-weighted harm coefficient, and empathy weighting; the deviation likelihood, deviation-preparation state, permitted deviation record, and dissonance buffer entry; the dependency graph and its reparation arcs; the integrity compliance score and mutation controller; and the deviation deductible, retention register, and aggregate retention.

This article discloses the structure and operation of those mechanisms as filed. It states no performance figure, deployment result, or benchmark, none of which the specification asserts. The declared parameters are named without asserted values, because the specification declares their form and bounds rather than particular settings, and the single numeric illustration above is the specification's own worked comparison of two harm coefficients. Nothing here characterizes any commonly owned application as prior art, and nothing asserts that any third-party system infringes any claim. The disclosure admits implementations developed after filing, alternative modulator and impact mappings, and additional need types, provided the permission semantics remain operative: that a deviation likelihood exceeding unity admits a forbidden mutation as a permitted deviation rather than a fault, that the admission is appended with the quantities that produced it, and that accumulated undischarged harm forecloses the permission before it compounds without bound.

Integrity Quantities and Permitted Deviation (/permitted-deviation)

[All 49 steps → \(/inventive-steps\)](#)

Bounded, accountable deviation from declared values — measured, not forbidden.

Provisional application

PRIMARY TECHNICAL DISCLOSURE

- [Permitted Deviation: Bounded, Recorded Departure From an Agent's Declared Values \(/articles/permitted-deviation/bounded-departure-from-declared-values\)](#)

SECONDARY TECHNICAL

- Declared cross-partition coupling
- Attack-conditional propagation suspension
- Global saturation-defeat remedy
- Typed, dimensionless conduct divergence measure
- Non-revisable retrieval of the policy in force
- Policy-difference set and policy-explained divergence
- Residual divergence and its confined consumption
- Cross-agent residual divergence
- Single-computation routing restriction and one-way direction
- Impairment terminal state and release
- Cross-agent coupling function
- Propagation halt on undischarged reparation arcs
- Entropy trajectory available to forecasting and to underwriting disclosure

TERMINOLOGY

- [aggregate retention \(/glossary#aggregate-retention\)](#).
- [deviation deductible \(/glossary#deviation-deductible\)](#).
- [deviation likelihood \(/glossary#deviation-likelihood\)](#).
- [entropy-weighted harm coefficient \(/glossary#entropy-weighted-harm-coefficient\)](#).
- [need quantity \(/glossary#need-quantity\)](#).
- [permitted deviation \(/glossary#permitted-deviation\)](#).
- [self-esteem aggregate \(/glossary#self-esteem-aggregate\)](#).

[Integrity Quantities and Permitted Deviation overview → \(/permitted-deviation\)](#).