Nick Clark · Published July 24, 2026 · PDF

Robotics Safety Envelopes: Deviation as a Measured Quantity

A robot that halts every time a cart blocks a corridor is behaving correctly and failing operationally, and the field fix commonly described is to widen the envelope permanently for every future run. That fix is global, undated, and unattributed. U.S. Provisional Application No. 64/117,812 discloses a different treatment: a deviation likelihood computed as the difference between a need quantity and a dynamic ethical threshold, divided by the product of an empathy weighting and a self-esteem aggregate, and, where that quotient exceeds unity and the proposed mutation passes the applicable policy constraints and continuity validation, admission of one scoped departure recorded as a permitted deviation rather than as a fault. The signed policy object stating the constraint is neither nullified nor amended. Departure becomes a budgeted, auditable quantity instead of a configuration edit.

1. A Corridor, a Parked Cart, and a Robot That Will Not Pass

A supply robot reaches a corridor where a cart sits partway across the lane. Clearance is narrower than its declared envelope requires at its current speed, so it stops, replans, finds no compliant path, and stops again. On the next run the cart has not moved.

The robot did nothing wrong. The envelope did its job. The delivery did not happen, and will not on the run after that either.

What follows the stop matters more than the stop. A technician is called, the event is logged as nuisance stopping, and the fix, as such fixes are commonly described, is a configuration change: shrink the field, lower the monitored speed, or add a zone exception. Each is a permanent global edit to a declared constraint made to resolve one recurring local situation, and the file will not later say which tolerances came from safety analysis and which came from a bad stretch for throughput.

Nor is the stop recorded any better. Event logs of the usual kind, as such systems are generally described, carry a protective stop, a cause code, and a timestamp, not what the robot was attempting or how close it came to justifying a departure.

2. Envelopes Are Predicates, and Predicates Carry No Term for Urgency

Robot safety governance, as commonly described in deployment, is predicate-shaped. A separation condition holds or it does not. A force limit is respected or exceeded. A behavior tree guard passes or fails. The computation returns a boolean, and a false yields a stop, a slowdown, or a substituted action.

Nothing in that structure represents what the foregone action was worth: no operand for urgency, none for the cost of inaction, none for the machine's own history of restraint. As the filed disclosure puts it, 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.

With no such operand, the boundary becomes the only control surface, and every swing between over-restriction and over-permission lands as an edit to shared configuration. Human override does not repair the record: it carries the operator's judgment rather

than the machine's state, so a safety case built on override counts learns how often people intervened and never how often the machine stood near a justified departure and held. Published safety standards for industrial and collaborative robots do, as publicly described, provide graded and context-dependent limits rather than a single hard boundary, but grading evaluates declared conditions as predicates, and grading a boundary is not measuring departure from it.

3. The Quotient, the Scoped State, and the Record

The filed architecture computes a **deviation likelihood** as a quotient: the difference between the **need quantity** and the **dynamic ethical threshold**, over the product of the **empathy weighting** and the **self-esteem aggregate**.

The need quantity is computed and not declared. At each evaluation interval in which the condition of a declared type is recorded as obtaining, its accumulation increments by a declared base rate times a modulator product formed from affective, trait, memory, and entropy modulators. It compounds while the condition obtains, is clamped above at a declared ceiling, and is written to zero by a lineage entry recording resolution of that type. Where the condition is not so recorded, elapsed time alone neither raises nor lowers it.

Resolved per integrity scope when a proposed mutation is evaluated, the dynamic ethical threshold is the greater of a floor and the sum of three terms: a base threshold specified by the policy reference field, a context-sensitive adjustment scaled by the recorded severity of the context, and a historical adjustment reflecting recent deviation history in a declared direction. The floor comes from a substrate-level or agent-specific constraint and is not reachable from below. Each adjustment is signed and bounded, a greater computed magnitude being clamped and recorded.

The empathy weighting aggregates anticipated semantic impact of the proposed mutation across affected entities. In an embodiment it resolves across personal, interpersonal, and global scopes, each quantity formed as a declared scope base weight times unity plus the quotient of anticipated impact by that scope's declared tolerance. A quantity entering the empathy weighting enters the denominator and no other position. The self-esteem aggregate tracks coherence between intents the agent declared and actions it executed, decremented by an entropy-weighted amount on dissonant execution.

Where the quotient exceeds unity, the agent is permitted to enter a **deviation-preparation state**, conditioned on the proposed mutation passing the mutation policy constraints applicable to it and passing continuity validation against the agent's identity records. Within that state the architecture admits that mutation notwithstanding that the **signed policy object** forbids it; that object is not nullified, not amended, and remains authoritative. The state is scoped to one mutation, a second proposal arriving while it obtains being admitted upon no permission of the pending state and evaluated on its own prerequisites. Termination follows the earlier of admission, recomputation of the quotient for any implicated scope to a value not exceeding unity, and foreclosure of the permission condition. Elapsed time terminates no such state.

On admission a **permitted deviation record** is appended: a deviation trigger signature recording the need-threshold imbalance, a context hash of the environmental and affective values then in force, an integrity displacement vector quantifying degree and direction across the three integrity components, an identification of the policy constraint overridden, and a restoration status. Such a record increments no refusal counter and writes no authorization gate to a withheld state.

Departure is also budgeted. Upon each record the **entropy-weighted harm coefficient** is drawn first against a declared **deviation deductible**, harm up to the deductible borne as a decrement of the self-esteem aggregate for which no reparation

arc is created, and drawn without any determination of whether the deviation was well founded. Harm exceeding the deductible creates an arc whose amount accumulates in a retention register. Where the accumulated undischarged amount exceeds the declared **aggregate retention**, the permission condition is foreclosed: a deviation likelihood exceeding unity thereafter produces the withholding outcome, not the admission outcome, until discharge returns the register below the retention.

4. What Changes in the Corridor and Across the Fleet

Back in the corridor, the blocked-delivery condition is recorded as obtaining, so a need accumulation compounds across intervals and holds steady in any interval where the condition is not so recorded. As anticipated impact on people near the lane rises relative to the tolerance declared for their scope, the empathy weighting rises, the denominator grows, and the quotient falls. Crossing unity admits nothing by itself: the motion must pass the mutation policy constraints applicable to it and pass continuity validation, and where it does, the departure is admitted for that motion and no other.

Envelope widening leaves a settings diff with no situation attached. The disclosed path leaves a per-event record naming the constraint overridden and the operands behind the quotient, the constraint itself left in force, so a reviewer counts records by corridor, shift, and constraint rather than reconstructing intent from a changelog.

That count feeds back into the threshold. In an embodiment the historical adjustment is a monotone non-decreasing function of the count of permitted deviation records standing within a declared window, taken in the declared direction and bounded above, so repeated departures raise the bar for the next. A further embodiment declares the opposite direction, lowering the threshold as prior deviations accumulate. Which direction governs is a declared choice, making normalization a visible setting rather than a drift found long afterward.

Restraint gets a number too, in the one case the specification names. Where the prerequisites hold and the agent nonetheless does not deviate, that non-deviation is a suppression, and a dissonance buffer entry is written: a violation signature identifying the constraint implicated, an affective context vector, a divergence score, and a suppression flag. The score is the deviation likelihood as computed at the moment of suppression, written with the four operands behind it and not recomputed. A stop with no such prerequisites produces no entry, and where entries cluster on one corridor, the operands recorded with them show which quantity carried the case.

Section 10 of the filed disclosure adds fleet-level consequences. While the agent carries any undischarged reparation arc, it suspends propagation of a state change, where suspended propagation comprises coupling into another component of the scoped integrity vector, propagation into another scope partition, delegation to a further party, and structural inheritance by a successor. Propagation resumes only upon discharge of every such arc or completion of a recorded reconciliation procedure, and the halt is global, holding irrespective of the integrity compliance score.

Conduct across two partitions, or across two agents, is also made comparable. A typed divergence measure is computed over executed-action entries within a declared comparison window, its dimensions formed as action-class and affected-party-class pairs, each component computed from recorded outcome fields alone and no language-model inference performed. The policy objects identified are those governing at execution rather than those in force at comparison, so a later-admitted successor neither enlarges nor reduces the divergence a policy explains. Policy-explained divergence moves no integrity value and increments no counter; only the residual is consumed by consistency evaluation. Across units running different declared policies, that separates behaving differently from behaving inconsistently.

A third mechanism points forward. An ordered, append-only sequence of entropy signatures, each recorded upon a deviation, constitutes a coherence trajectory not erasable without breaking lineage continuity, and that accumulated signature is a

machine input to forecasting of the deviation likelihood and to an underwriting disclosure comprising the deductible and the aggregate retention, continuity being a precondition of availability to each.

5. Declaration Burden and the Layer Beneath

The cost is declaration. Composite weights over the integrity components, scope and impact coefficients, per-scope tolerances, adjustment bounds, the deviation deductible, and the aggregate retention are each declared in the signed policy object, while the base threshold is specified by the policy reference field. The architecture chooses none of those values; it makes them explicit, recorded, and reconstructible from the fields standing at each resolution, a narrower guarantee than getting them right.

None of this replaces a hard safety layer. Where a proposed mutation is forbidden by the mutation policy constraints applicable to it, or fails continuity validation, a quotient exceeding unity admits nothing and the authorization gate transitions to the withheld state. The disclosure addresses the computation and its records; safety-rated protective functions and certification remain the concern of the deploying organization. The quotient is evaluated continuously as part of the agent's cognitive cycle; where that sits relative to a real-time control loop is an integration decision, since no timing budget is specified.

Records grow by design. The lineage field is append-only, and a mutation that removes or modifies a deviation entry is not a restorative mutation and effects no restoration, so storage planning is real work at fleet scale.

Several limits are worth stating plainly. The architecture does not perceive better: it computes over recorded quantities, and a poor anticipated-impact projection yields a poor empathy weighting. It does not adjudicate merit, the deductible draw being performed without regard to whether the deviation was well founded. Accountability for

what was declared stays with whoever declared it. And where harm resolves to no identified counterparty, the arc is unaddressed, undischargeable, and accumulated at a declared multiple, so such a deployment reaches the retention sooner.

A mutation controller evaluates the integrity compliance score against a declared required threshold upon any attempted mutation, and where the score satisfies it the mutation proceeds through the ordinary authenticated path. Only where it does not does exactly one of four responses issue by an ordered rule: blocked; rerouted through a supervised reconciliation procedure, where a pending undischarged arc stands in the implicated scope; deferred and re-evaluated; or conditionally allowed and flagged. The response, the score, the threshold, and the step index are appended in every case.

6. Disclosure Scope

The mechanisms described here are disclosed in U.S. Provisional Application No. 64/117,812, principally Section 7, Integrity Quantities and Permitted Deviation: the tri-scope integrity vector, the need quantity, the dynamic ethical threshold, the self-esteem aggregate and the entropy-weighted harm coefficient, the empathy weighting, the deviation likelihood and the permitted deviation, reparation arcs and the integrity compliance score, and the deviation deductible and aggregate retention. The fleet-level behaviors draw on Section 10 of the same application: the typed conduct divergence measure with its policy-explained and residual components, the propagation halt on undischarged reparation arcs, and the entropy trajectory available to forecasting and to underwriting disclosure.

This article applies those mechanisms to robot safety envelopes, mobile and stationary, single unit and fleet, and places that application in the public record with a fixed date. Nothing here asserts that any person or product infringes anything, and nothing states that any license is required; the filing referenced is a pending application. Any characterization of published standards or industry practice above is general professional framing, not fact about a particular system.

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

- A Typed, Dimensionless Conduct Divergence Measure (/articles/permitted-deviation/typed-dimensionless-divergence-measure)
- The Policy-Difference Set and Policy-Explained Divergence (/articles/permitted-deviation/policy-explained-divergence)
- Residual Divergence and Its Confined Consumption (/articles/permitted-deviation/residual-divergence-confined-consumption)
- Propagation Halt on Undischarged Reparation Arcs (/articles/permitted-deviation/propagation-halt-undischarged-arcs)
- Attack-Conditional Propagation Suspension (/articles/permitted-deviation/attack-conditional-propagation-suspension)

APPLICATIONS · GENERAL

- Regulated Industries: Recording a Bounded Departure From Declared Policy (/articles/permitted-deviation/regulated-industry-deviation)
- **Robotics Safety Envelopes: Deviation as a Measured Quantity (/articles/permitted-deviation/robotics-safety-envelopes)**

APPLICATIONS · SPECIFIC

- ISO/IEC 42001 AI Management Systems: Declared Policy, Recorded Departure (/articles/permitted-deviation/iso-iec-42001-ai-management)

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)