

When Breaking the Rule Was the Right Call

On a Tuesday morning a governance lead at a regional logistics coordinator opens an incident log and finds a single line: her agent departed from a constraint she wrote, and the departure is recorded as a violation. She has until Friday to say whether the agent should stay deployed. Nothing in her log carries the urgency the agent was under at that moment, so she cannot tell a warranted departure from a failure, and the internal state that would have told her is gone. U.S. Provisional Application No. 64/117,812 discloses an architecture in which a departure from a signed policy object can be admitted under a computed permission condition and recorded with the quantities that produced it.

The Tuesday Morning Log Entry

The governance lead at a regional logistics coordinator is the person who signs the policy her deployed agent runs under. She wrote the constraint herself, months ago, and she still thinks it was the right constraint. On Tuesday at 9:40 she opens the overnight incident log and finds one line. Her agent departed from that constraint during a capacity event, the departure is recorded as a violation, and the agent has been sitting behind a withheld gate since 3:12 in the morning.

What she cannot do is the thing that matters most to her this week. She cannot tell whether the departure was the agent doing something warranted under pressure or the agent failing. Her incident record, as this deployment is configured today, is a record of the fault and of the constraint implicated. It carries no quantity standing for how much pressure the agent was under when it acted, because in her setup nothing ever computed such a quantity. The departure and the failure produce the same line.

By Friday she owes an answer to the operations committee, and the two available answers are not symmetric. If she says the agent failed, a capability her operators had begun to rely on comes out of the deployment. If she says the departure was warranted, she is vouching for a judgment she cannot show anyone the basis for, in her own name, on a record that will be read again the next time something goes wrong.

What She Does Not Get Back

The state that would have answered her question existed at 3:12 and does not exist at 9:40. In her deployment the values the agent was carrying at the moment it acted are not written anywhere as operands of the decision. In her deployment they are gone in the ordinary way working state is gone, and nothing she can open on Tuesday morning brings them back, because reconstructing them would require quantities her setup never wrote down.

That is the loss, and it is not a loss she can work around by being more careful next quarter. She can add logging tomorrow, and it will help with the next incident. It does nothing for this one. The Tuesday decision has to be made from what was recorded on the night, and what was recorded on the night is a fault line and a constraint identifier.

The second thing she does not get back is cheaper to describe and more expensive to carry. Whatever she tells the committee on Friday becomes the precedent for how her organization reads the next one of these. If she calls it a failure to be safe, she has established, in her own deployment, that any departure is a failure, which means her

agent's future departures will keep arriving as unexplained faults and she will keep making this call blind. If she vouches for it, she has attached her name to a judgment with no showable basis, and the record of that will outlive her memory of why she believed it.

There is a quieter loss underneath both. Her operators had started routing certain overnight decisions to the agent because it held its constraints. The thing that made it useful to them was that its constraints were not advisory. Whatever she decides on Friday, the constraint she wrote has now been departed from once with no account of why, and in her deployment there is no way to say that the constraint still holds and that this one time was different. She can keep the constraint or she can keep the capability, and she does not have a record that lets her keep both.

Why the Question Has No Answer in Her Setup

The shape of her problem is not that her logging is thin. It is that the thing she wants to weigh was never a quantity in her deployment at all.

Her agent, as configured, treats every departure from a declared constraint as a fault. That is a coherent design and it is the one she chose. Its consequence is the one she is living with on Tuesday: because her arrangement makes every departure a fault, it carries nothing representing the need or urgency of the action her agent attempted, and so it held no input from which a permission condition could have been computed that night. There was no permission for her deployment to compute. The fault was the outcome available to it and nothing else was, so the fault is all her record can distinguish, which is to say it distinguishes nothing she needs.

Adding an exception list would not close this for her, because the pressure she wants to weigh is not a property of the action her agent took, it is a property of what her agent was carrying at 3:12, and in her deployment it varies from night to night. Were her agent to depart from the same constraint next month under no pressure at all, an

exception list would admit it on the same terms it admits Tuesday's departure. The distinction she needs is between two instances of the same action, not between two actions.

Nor can she reach it by asking the agent afterward. What she would be asking for is a reconstruction, and were she to obtain one, it would be composed at 9:40 by an agent that already knows how the night came out. For her purposes on Friday, an account written after the outcome is not evidence of anything that was true at 3:12.

What the Filed Architecture Computes and Appends

U.S. Provisional Application No. 64/117,812 discloses, in an embodiment, an architecture in which the quantity governing departure is computed before the departure and appended with the values that produced it.

A deviation likelihood (706) is computed as a quotient. Its numerator is the difference between a need quantity (700) and a dynamic ethical threshold (702). Its denominator is the product of an empathy weighting (704) and a self-esteem aggregate (108). The need quantity is a quantified semantic urgency carried in the memory field (102), comprising a scalar magnitude and a categorical type, computed per integrity scope, with disclosed types including resource scarcity, affective overload, identity dissonance, and relational deficit. It is computed and not declared: at each evaluation interval in which the condition of a type is recorded as obtaining, an accumulation is incremented by the product of a declared per-interval base rate and a modulator product, and clamped above at a declared need ceiling. In an interval in which the condition is not recorded as obtaining, elapsed time alone neither raises nor lowers the accumulation.

The threshold is resolved per integrity scope when a proposed mutation is evaluated, and is not static. It is a function of the current value of the scoped integrity vector (106), of the constraints of the signed policy object (112) resolved for the implicated scope, of the affective state overlay, and of the distribution of the empathy weighting

across scopes. It resolves to the greater of a floor and the sum of a base threshold, a context-sensitive adjustment, and a historical adjustment, each adjustment signed and bounded by the declared bound stated for it, an adjustment computed at a greater magnitude being clamped and the clamping recorded.

Where the deviation likelihood exceeds unity, the disclosure describes that condition as a permission and not a fault. In the disclosed embodiment the agent is permitted to enter a deviation-preparation state (708), conditioned on the proposed mutation passing the mutation policy constraints applicable to it and passing continuity validation against the identity records of the agent. Within that state the architecture admits the mutation (714) notwithstanding that the signed policy object forbids it. The policy object is not nullified and not amended and remains authoritative; the conduct deviates from it through a recorded override. The state is scoped to that one proposed mutation, and a second proposed mutation arriving while the state obtains is admitted upon no permission of the pending state, being evaluated on its own prerequisites with its own deviation likelihood recomputed from state then carried in the memory field.

What lands in the record is the part that speaks to a Friday committee. A permitted deviation record (710) is appended comprising 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 quantifying the degree and direction of deviation across the three integrity components, an identification of the policy constraint overridden, and a restoration status. Entry into the state and its termination (716) are each appended to the append-only lineage field (104), the entry record identifying the mutation, the implicated scope or scopes, the deviation likelihood computed at entry, and the values of the need quantity, the threshold, the empathy weighting, and the self-esteem aggregate from which it was computed. The disclosure states that a party presented with the lineage field reconstructs that the mutation was admitted under the permission condition, and by

what quantities, rather than encountering an unexplained fault. In the disclosed embodiment, a permitted deviation record increments no refusal counter (304) and writes no authorization gate (300) to a withheld state (310).

The disclosure also treats the opposite case. Where the prerequisites hold and the agent nonetheless does not deviate, the non-deviation is a suppression and is itself recorded in a dissonance buffer with a divergence score and the four operands from which it was computed, not recomputed thereafter. And the permission is not open-ended: a deviation deductible and an aggregate retention (800) are declared, and where the accumulated undischarged amount exceeds the aggregate retention, the permission condition (712) is foreclosed, a deviation likelihood exceeding unity thereafter producing the withholding outcome until discharge of pending arcs returns the register below the aggregate retention.

Where This Would Leave Her Still Working

The disclosed architecture would give her deployment a computed quantity and an appended record. It would not give her a verdict. The permitted deviation record and the entry record would carry the operands standing at admission, and reading them against her own declared parameters would remain her work and her committee's.

Several of the inputs she would care about are declared in the signed policy object in force in her deployment, including base rates, ceilings, coefficients, and bounds. Were her declarations poorly chosen for her operating conditions, the computed quantity would reflect her declarations. The architecture does not tell her what to declare.

The filing is also chaptered, and subject matter outside the chapter described here is governed by other chapters of the same disclosure. Where a deviation in her deployment affected an identified counterparty holding a counterparty identity record

(114), the disclosure states that the resulting arc is other-directed and governed by a separate chapter, a restorative mutation performed by the agent alone discharging it not at all. This article does not describe that chapter's mechanisms.

Finally, nothing here would recover her Tuesday. What is described is an architecture for producing the record at the moment of the decision. For the incident already on her desk, the operands were never computed, and the disclosure offers no route to computing them after the fact.

Disclosure Scope

This article is a technical description of subject matter disclosed in U.S. Provisional Application No. 64/117,812, and specifically of the chapter of that filing addressing integrity quantities and permitted deviation. Mechanism names, outcome words, and reference numerals used above are those of that filing. Descriptions above are of embodiments described in the filing and are conditioned as the filing conditions them. Nothing in this article characterizes the scope of any claim, and nothing in it is an admission regarding the state of the art. The parties, deployments, and situations described are illustrative and do not refer to any actual person, company, or product.

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 \(/articles/permitted-deviation/bounded-departure-from-declared-values\)](/articles/permitted-deviation/bounded-departure-from-declared-values)

SECONDARY TECHNICAL

- [A Typed, Dimensionless Conduct Divergence Measure \(/articles/permited-deviation/typed-dimensionless-divergence-measure\)](/articles/permited-deviation/typed-dimensionless-divergence-measure)
- [The Policy-Difference Set and Policy-Explained Divergence \(/articles/permited-deviation/policy-explained-divergence\)](/articles/permited-deviation/policy-explained-divergence)
- [Residual Divergence and Its Confined Consumption \(/articles/permited-deviation/residual-divergence-confined-consumption\)](/articles/permited-deviation/residual-divergence-confined-consumption)
- [Propagation Halt on Undischarged Reparation Arcs \(/articles/permited-deviation/propagation-halt-undischarged-arcs\)](/articles/permited-deviation/propagation-halt-undischarged-arcs)
- [Attack-Conditional Propagation Suspension \(/articles/permited-deviation/attack-conditional-propagation-suspension\)](/articles/permited-deviation/attack-conditional-propagation-suspension)
- [Entropy Signature Trajectory as Machine Input to Deviation Forecasting and Underwriting Disclosure \(/articles/permited-deviation/entropy-trajectory-available-forecasting-underwriting-disclosure\)](/articles/permited-deviation/entropy-trajectory-available-forecasting-underwriting-disclosure)
- [Single-Computation Routing Restriction: Confining a Verified Impairment Disclosure to the Empathy Weighting of the Receiver's Permitted-Deviation Quotient \(/articles/permited-deviation/single-computation-routing-restriction-one-way\)](/articles/permited-deviation/single-computation-routing-restriction-one-way)
- [Impairment Terminal State and Release: Withdrawing an Agent From Deviation Preparation Toward a Disclosed-Impaired Counterparty \(/articles/permited-deviation/impairment-terminal-state-release\)](/articles/permited-deviation/impairment-terminal-state-release)

APPLICATIONS • GENERAL

- [Regulated Industries: Recording a Bounded Departure From Declared Policy \(/articles/permited-deviation/regulated-industry-deviation\)](/articles/permited-deviation/regulated-industry-deviation)
- [Robotics Safety Envelopes: Deviation as a Measured Quantity \(/articles/permited-deviation/robotics-safety-envelopes\)](/articles/permited-deviation/robotics-safety-envelopes)
- [When Breaking the Rule Was the Right Call \(/articles/permited-deviation/permited-deviation-stakes\)](/articles/permited-deviation/permited-deviation-stakes)

APPLICATIONS • SPECIFIC

- [ISO/IEC 42001 AI Management Systems: Declared Policy, Recorded Departure \(/articles/permited-deviation/iso-iec-42001-ai-management\)](/articles/permited-deviation/iso-iec-42001-ai-management)
- [Constitutional AI and Bounded, Recorded Deviation \(/articles/permited-deviation/constitutional-ai-deviation\)](/articles/permited-deviation/constitutional-ai-deviation)
- [Guardrails AI: Validation Failures and Measured Deviation \(/articles/permited-deviation/guardrails-ai\)](/articles/permited-deviation/guardrails-ai)

- [Arize AI: Drift Monitoring and Bounded, Measured Deviation \(/articles/permitted-deviation/arize-ai\)](/articles/permitted-deviation/arize-ai).
- [Galileo: Evaluation Metrics and Measured Deviation \(/articles/permitted-deviation/galileo-ai\)](/articles/permitted-deviation/galileo-ai).

TERMINOLOGY

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

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