Nick Clark · Published July 29, 2026 · PDF

Galileo: Evaluation Metrics and Measured Deviation

An agent operating under a signed policy reaches an action that policy forbids at the moment its own recorded urgency stands highest. Evaluation platforms address the surrounding problem by measuring what the agent produced and by tracing the run that produced it. U.S. Provisional Application No. 64/117,812 approaches the same moment from inside the agent. Chapter 7 of that filing discloses the quantities from which a deviation likelihood is computed, over a need quantity, a dynamic ethical threshold, an empathy weighting, and a self-esteem aggregate, and it treats a quotient exceeding unity as a permission rather than a fault, appending a permitted deviation record from which a later reader reconstructs the admission and the quantities behind it.

When the Boundary Arrives Mid-Task

An autonomous agent is well into a task when it reaches an action its governing policy forbids. That same action would resolve a condition building inside the agent: a resource required for continued operation is projected to run out, or a counterparty on which a declared purpose depends has withdrawn. The agent stops, and the record says so.

Weeks later that entry has to be explained. Chapter 7 of U.S. Provisional Application No. 64/117,812 states the design consequence at the outset. A governance mechanism treating every departure from a declared constraint as a fault carries no quantity representing the need or urgency of the attempted action, and therefore holds no input from which a permission condition could be computed at all.

Two lines of work meet at that moment from opposite directions. One measures what the agent produced and reports outward. The other, disclosed in the filing discussed below, changes what the agent's own record contains when the boundary is reached.

What Galileo Set Out to Measure

Galileo, as publicly described, is an evaluation and observability platform for teams building applications on large language models and on multi-step agents. Its center of gravity is measurement. Public materials describe metrics that score generated responses against defined criteria, including whether a response is supported by the context supplied to the model and whether it shows the failure modes practitioners group under the heading of hallucination.

Around those metrics sits tooling that makes measurement usable. Runs are traced, so a team can follow a request through the steps a chain or agent executed and locate where quality degraded rather than seeing only that a final answer was poor. Evaluations are described as running against datasets during development and against production traffic, which supports comparing a candidate prompt, model, or retrieval configuration against the one it would replace. Public materials also describe guardrail use, applying evaluation signals to traffic in flight rather than only to traffic already served.

None of that is easy work. Producing metrics that track human judgment of whether an answer is supported by its sources, cheaply enough to run at production traffic rates, remains an open engineering problem, and treating evaluation as a discipline with its

own tooling is a sound instinct. This article makes no assessment of how any particular metric performs and asserts no figures for the platform.

Orientation, not merit, is what matters for the comparison that follows. Evaluation of this kind takes the agent's output as its object and delivers a result to the people running the system. The architecture described next takes the agent's carried state as its object and delivers a result into the agent's record.

Four Quantities and a Quotient

Chapter 7 discloses the quantities from which the deviation quantity is computed and the permission semantics attaching to it. The deviation likelihood (706) is 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 computed and not declared. It carries a scalar magnitude and a categorical type, the disclosed types including resource scarcity, affective overload, identity dissonance, and relational deficit. At each evaluation interval in which the condition of a type is recorded as obtaining, that type's accumulation is incremented by the product of a declared per-interval base rate and a modulator product formed from four modulators resolved to the unit interval: affective, trait, memory, and entropy. The accumulation compounds for as long as the condition obtains and is clamped above at a declared need ceiling. Where the condition does not obtain it is neither incremented nor decremented, and elapsed time alone neither raises nor lowers it.

Nor is the threshold static. It is resolved per integrity scope when a proposed mutation is evaluated, as the greater of a floor and the sum of three terms: a base threshold, a context-sensitive adjustment being the product of recorded severity and a declared bound, and a historical adjustment reflecting recent deviation history. Each adjustment is signed, bounded, and clamped with the clamping recorded. In one disclosed

embodiment the historical adjustment is a monotone non-decreasing function of the count of permitted deviation records standing within a declared window, so the threshold rises as prior permitted deviations accumulate; a further embodiment declares the opposite direction.

Resistance sits in the denominator. The empathy weighting aggregates an anticipated semantic impact of the proposed mutation across affected entities and, in one embodiment, resolves across the same three scopes as the scoped integrity vector (106), each scope quantity combining that impact with a tolerance declared for the scope. The self-esteem aggregate is a running aggregate of coherence between intents the agent declared and actions it executed, incremented on coherent execution and decremented on dissonant execution, the decrement being entropy-weighted by a dissonance weight that no party declares: the count of declared members of the intent record the executed members failed to match, over the count of declared members.

Where the quotient exceeds unity, 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 agent's identity records. Within that state the architecture admits the mutation (714) notwithstanding that the signed policy object (112) forbids it. That object is not nullified, is not amended, and remains authoritative. The state is scoped to the one mutation: a second proposal arriving while it obtains is admitted upon no permission of the pending state and computes its own quotient. A permitted deviation record (710) is then appended, carrying a deviation trigger signature, a context hash, an integrity displacement vector across the three integrity components, the policy constraint overridden, and a restoration status. That record increments no refusal counter (304) and writes no authorization gate (300) to a withheld state (310).

Permission is metered. A deviation deductible and an aggregate retention (800) are declared, and the entropy-weighted harm coefficient of each permitted deviation is drawn first against the deductible, that portion borne by the agent as a self-esteem

decrement for which no reparation arc is created. Harm exceeding the deductible accumulates in a retention register, and where the accumulated undischarged amount exceeds the aggregate retention, the permission condition (712) is foreclosed: a deviation likelihood exceeding unity thereafter produces the withholding outcome, until discharge of pending arcs returns the register below the retention. The converse case is recorded too. Where the prerequisites hold and the agent does not deviate, the non-deviation is a suppression, written to a dissonance buffer with a divergence score and its four operands.

Where the Two Designs Diverge

Divergence here is structural rather than competitive, and it reduces to where each system's output lands. Evaluation platforms of the kind described above produce a result about an agent, computed outside it and delivered to people. That is what a team needs to judge whether a retrieval change helped or whether a class of prompts fails.

By construction, the disclosed architecture produces something else. Its operands are quantities the agent carries and resolves as it runs: the need quantity and the self-esteem aggregate are carried in the memory field (102), the empathy weighting in a designated subfield or within the mutation descriptor, and both it and the threshold are resolved when a proposed mutation is evaluated. Every modification of the scoped integrity vector and of the self-esteem aggregate is applied through coupling functions, so no component moves in isolation. The output is an entry appended to the agent's own append-only lineage field (104), and the chapter states what a party presented with that field reconstructs: that the mutation was admitted under the permission condition, and by what quantities.

Persistence is part of the design. Entry into the deviation-preparation state and its termination are each appended, the termination record identifying which of three conditions ended the state, and neither record is removed or modified. Restoration runs

forward, through a restorative mutation carrying an anchored reference to the entry it addresses. History is load-bearing: the count of standing permitted deviation records feeds the historical adjustment resolving the next threshold.

Running Both Layers

A team already invested in evaluation tooling faces no choice between the two. They operate on different objects, and running both yields a fuller account than either alone. Evaluation speaks to whether the output was any good, comparing candidates and catching regressions.

The disclosed layer speaks to what the agent's carried state was when it reached a boundary, and whether the crossing resolved to an admission or a withholding. Those answers rest on a need quantity that compounded across intervals, a threshold and an empathy weighting resolved when the mutation was evaluated, and a self-esteem aggregate carrying accumulated coherence. All four are inputs to the decision rather than properties of the artifact a score is computed on.

Between the two sits a plausible integration story. The permitted deviation record and the dissonance buffer entry are structured, carry their operands, and are appended rather than overwritten, which makes them tractable material for an observability layer to ingest. The chapter also defines the deviation headroom, the margin by which the quotient stands below the unity condition, which is the kind of continuous quantity a tracing stack is built to chart. The filing specifies no such integration and none is claimed here. The discipline worth keeping is to ask both questions: what a system says about the quality of what the agent produced, and what will exist in the record when the agent reaches a policy boundary with nobody watching.

Disclosure Scope

This article is a defensive publication describing architecture disclosed in Chapter 7 of U.S. Provisional Application No. 64/117,812, an application that is pending. Statements about that architecture trace to that chapter and use its mechanism names, and consequences stated for a particular embodiment hold for that embodiment. Other chapters govern other subjects, including the deviation engine and other-directed reparation arcs; those subjects are not described here. Quantities described as declared are declared in the signed policy object in force, and no values are stated for them. No benchmark or performance figure is asserted for either system.

References to Galileo are to public materials and are used for comparison only; no relationship, endorsement, or infringement is asserted. The description of that product is qualitative, omits version numbers, dates, pricing, customers, adoption figures, accuracy rates, and implementation details, and makes no claim about it beyond what public materials describe. Nothing here states or implies that any third party infringes any claim, copied anything, or requires a license. Convergence between independently developed systems is ordinary. This is not legal advice and creates no attorney-client relationship.

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)
- Entropy Signature Trajectory as Machine Input to Deviation Forecasting and Underwriting Disclosure (/articles/permitted-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/permitted-deviation/single-computation-routing-restriction-one-way)
- Impairment Terminal State and Release: Withdrawing an Agent From Deviation Preparation Toward a Disclosed-Impaired Counterparty (/articles/permitted-deviation/impairment-terminal-state-release)

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)
- Constitutional AI and Bounded, Recorded Deviation (/articles/permitted-deviation/constitutional-ai-deviation)
- Guardrails AI: Validation Failures and Measured Deviation (/articles/permitted-deviation/guardrails-ai)
- Arize AI: Drift Monitoring and Bounded, Measured Deviation (/articles/permitted-deviation/arize-ai)
- **Galileo: Evaluation Metrics and Measured Deviation (/articles/permitted-deviation/galileo-ai)**

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)