Nick Clark · Published July 29, 2026 · PDF

Constitutional AI and Bounded, Recorded Deviation

Constitutional AI, as publicly described, shapes a model's behavior by training it against an explicit written set of principles, so that refusal and compliance emerge from learned dispositions. Adaptive Query's filings address a different layer of the same problem: what a governed agent's runtime does in the moment a declared constraint and a recorded urgency point in opposite directions. U.S. Provisional Application No. 64/117,812 discloses, in an embodiment, a deviation likelihood computed from a need quantity, a dynamic ethical threshold, an empathy weighting, and a self-esteem aggregate, and treats a quotient exceeding unity as a permission rather than a fault, admitting the forbidden mutation only where stated conditions are satisfied and appending a permitted deviation record that identifies the constraint overridden.

When the rule holds and the situation does not

A governed agent reaches a case where a declared constraint and the agent's own recorded state pull in opposite directions. Its policy forbids a class of action. Its state records a mounting reason to take that action anyway: a resource required for continued operation is nearing exhaustion, a contradiction between declared roles has been recorded, a counterparty on which a declared purpose depends has withdrawn.

Two responses are commonly available. The agent refuses, and the refusal is stored as a category label that carries none of the quantities behind it. Or the agent proceeds, and the departure is afterward hard to distinguish from a defect.

Neither response leaves behind what a reviewer actually needs, which is a record of what was weighed, how close the call was, and what remains owed. The architecture disclosed in U.S. Provisional Application No. 64/117,812 puts the point structurally: 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. The filing's answer is to carry that quantity and to compute against it.

Training against a written standard

Constitutional AI, as publicly described by Anthropic, trains a model against an explicit written set of principles rather than relying solely on case-by-case human labeling of preferred outputs. As publicly described, the model critiques and revises its own responses against those principles, and feedback derived from that process is used in training. The set of principles is itself published text, which means the standard a model is being held to can be read, argued with, and revised by people who are not inside the training loop.

That is a substantial contribution to a real problem. Behavior learned from an articulated standard is legible in a way that behavior learned from aggregate preference signals is not: a reviewer can point at a principle and ask whether it was applied correctly. Written principles can also be extended and amended without re-collecting the underlying human judgments, and the same document can inform evaluation as well as training.

The category served here is behavioral alignment, bringing a system's dispositions into line with a stated standard across the enormous space of situations nobody enumerated in advance. The architecture disclosed in the filing takes a standard as given and addresses a different question: what a runtime does with one specific proposed act, evaluated against quantities carried in the agent's own memory at that moment. Both concerns can be present in a single deployed system, and neither substitutes for the other.

Inside the filed mechanism: a quotient that can exceed unity

In the disclosed embodiment, the deviation likelihood (706) is a quotient. Its numerator is the difference between the need quantity (700) and the dynamic ethical threshold (702). Its denominator is the product of the empathy weighting (704) and the self-esteem aggregate (108). Each operand is a quantity carried in the memory field (102) and reconstructible from the agent's records.

Need is computed, not declared. Per type and per integrity scope, an accumulation is incremented at each evaluation interval in which the condition of that type is recorded as obtaining, by the product of a declared base rate and a product of four modulators resolved from the affective state overlay, from persistent trait parameters, from prior unresolved occurrences of the same type within a declared window, and from recorded entropy exposure. The named types comprise resource scarcity, affective overload, identity dissonance, and relational deficit. The accumulation compounds for so long as the condition obtains and is clamped above at a declared ceiling. In an interval where the condition is not recorded as obtaining, the accumulation is neither incremented nor decremented, elapsed time alone neither raising nor lowering it. Where an entry of the append-only lineage field (104) records the resolution of a type, that accumulation is written to zero, and the entry identifies the type, the scope, and the amount extinguished.

Thresholds move too. The dynamic ethical threshold is resolved per scope as the greater of a floor and the sum of three terms: a base threshold, a context-sensitive adjustment, and a historical adjustment reflecting recent deviation history. Each adjustment is signed and bounded by a declared magnitude, and one computed at a greater magnitude is clamped, with the clamping recorded. The direction of the historical adjustment is itself declared: in one embodiment the threshold rises as prior permitted deviations accumulate, and in a further embodiment it falls, reflecting normalized deviation patterns. The denominator carries the agent's resistance. The empathy weighting aggregates an anticipated semantic impact of the proposed mutation across affected entities, and the self-esteem aggregate tracks coherence between intents the agent declared and actions it executed, decremented by an entropy-weighted amount where the two diverge.

Where the quotient exceeds unity, the agent is permitted to enter a deviation-preparation state (708). That permission is conditioned: the proposed mutation must pass the mutation policy constraints applicable to it and pass continuity validation against the identity records of the agent. Within that state the architecture admits the mutation notwithstanding that the signed policy object (112) forbids it. The policy object is not nullified and not amended, and it remains authoritative. The state is scoped to that one mutation, and a second proposal arriving while the state obtains is admitted upon no permission of the pending state; its own quotient is recomputed from state then carried.

A permitted deviation record (710) is then appended, comprising a deviation trigger signature recording the need-threshold imbalance at admission, a context hash of 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. It increments no refusal counter (304) and writes no authorization gate (300) to a withheld state (310). Reading the lineage field, a party reconstructs that the mutation was admitted under the permission condition, and by what quantities, rather than encountering an unexplained fault.

Restraint is recorded on the same terms. Where the prerequisites hold and the agent nonetheless does not deviate, the non-deviation is a suppression, and a dissonance buffer entry is written carrying a violation signature identifying the constraint implicated, an affective context vector, a suppression flag, and a divergence score together with the four operands from which it was computed.

Different layer, different failure mode

Divergence here is architectural rather than competitive, and three properties of the disclosed architecture are structural requirements of the filing rather than tuning choices.

Permission is per-mutation and non-transferable. Nothing about one admission licenses the next. The deviation-preparation state terminates upon the earliest of admission and appending of the record, recomputation of the quotient for any implicated scope to a value not exceeding unity, and foreclosure of the permission condition; entry and termination are each appended, and elapsed time terminates no such state.

Admission is not free. Under Section 7.9 the entropy-weighted harm coefficient of the deviation is drawn first against a declared deviation deductible, and harm up to the deductible is borne by the agent as a decrement of the self-esteem aggregate for which no reparation arc is created and no discharge is available. That drawing is performed without regard to, and without the agent performing, any determination of whether the deviation was well founded.

Permission can also run out. Where the accumulated undischarged amount exceeds the declared aggregate retention (800), the permission condition is foreclosed, and a deviation likelihood exceeding unity thereafter produces the withholding outcome instead of the admission outcome, until discharge of pending arcs returns the register below the retention.

What is owed is directed by whom the harm fell on. Where the affected-party class resolves to an identified counterparty holding a counterparty identity record (114), the arc is other-directed and, under the filing, governed by a separate chapter; a restorative mutation performed by the agent alone discharges it not at all. Where it resolves to no identified counterparty, the arc is unaddressed and undischargeable, accumulating in the retention register at a declared multiple. Elapsed time discharges no arc.

Coexistence, and the boundary of the claim

Composition, not competition, describes how these fit in a deployment. A model whose dispositions are shaped by an articulated set of written principles is a good candidate for this architecture, because the quality of the standard being deviated from determines what a deviation record means. The written principles supply the constraint; the signed policy object carries a constraint in a form the runtime can evaluate a proposed mutation against; the permitted deviation record captures the individual episodes in which the two pulled apart.

The boundary is worth stating plainly. The disclosed architecture does not make an underlying model behave well, and it holds no view on how a constraint was learned. It does not decide whether a given deviation was correct, since the deductible is drawn without any such determination. Its quantities are policy-declared, and the filing states no numeric values for the base rates, ceilings, coefficients, deductible, or aggregate retention, so a deployment that declares them badly gets a badly calibrated agent together with an accurate record of its own miscalibration. The mechanism further assumes an append-only lineage field that survives the agent, which is an operational commitment rather than a property the arithmetic confers.

Disclosure Scope

This article describes subject matter disclosed in U.S. Provisional Application No. 64/117,812, a pending application. Mechanism names, reference numerals, and outcome conditions are taken from that filing and are described as disclosed in an embodiment. Quantities described as declared are declared in the signed policy object in force; the filing states no numeric values for them, and none should be inferred here. Nothing in this article is a claim construction, an opinion of counsel, or a representation about the scope of any patent that may issue.

References to Constitutional AI are to public materials and are used for comparison only; no relationship, endorsement, or infringement is asserted.

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)

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)