Nick Clark · Published July 24, 2026 · PDF

ISO/IEC 42001 AI Management Systems: Declared Policy, Recorded Departure

An AI management system asks an organization to declare a policy and then show that its AI conforms to it. Autonomous agents make that showing hard to produce. A refusal writes nothing, a break-glass override writes prose, and conduct between agents belonging to different organizations lands outside the boundary any single management system covers. The architecture disclosed in U.S. Provisional Application No. 64/117,812 computes a deviation likelihood 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. Where that quotient exceeds unity, and where the proposed mutation passes the mutation policy constraints applicable to it and passes continuity validation, the departure is admitted as a permitted deviation and recorded with the quantities behind it, the signed policy object remaining authoritative and unamended.

The Evidence That Never Gets Written

A supplier's agent commits to terms its own declared policy does not permit. The counterparty's agent, run by another company under another policy, sees the commitment and not the reasoning. Asked later to show what happened, the supplier's

compliance function finds either nothing, because the action was refused inside the governed system and taken outside it, or a log entry noting an override with free text attached.

Management system standards for AI rest on a straightforward bargain. Declare a policy, and be able to show conformance to it, including where conformance failed and what followed. That bargain assumes the evidence exists, which it usually does where people follow written procedures. Where agents act autonomously, at machine frequency, against agents belonging to other organizations, it often does not.

Three kinds of evidence go missing. Restraint is one: where the conditions for a departure held and the system held anyway, ordinary logging records nothing, because from the log's point of view nothing happened. Departure itself is another, since a break-glass path returns an attestation rather than a measurement, and two agents facing identical conditions under one policy emit different sentences that cannot be compared or summed.

Hardest is the cross-organization case. A management system is scoped to the organization that holds it. When one company's agent departs from policy in a way that lands on another company's agent, the internal nonconformity record and the corrective action that follows sit behind a fence the affected party cannot see over, and neither is addressed to it.

What ISO/IEC 42001 Sets Out to Do

The account below is drawn from public materials describing the standard, kept to the level at which those materials are stable.

ISO/IEC 42001 is an international standard for an artificial intelligence management system, published jointly by ISO and IEC. It belongs to the same family as ISO's other management system standards and follows the harmonized structure they share. An

organization establishes its context and scope, secures leadership commitment, sets an AI policy and objectives, assigns roles, identifies and treats risk, operates under documented controls, monitors and measures, audits itself, reviews at the management level, and handles nonconformity with corrective action. The standard is certifiable, and accredited bodies assess organizations against it.

Subject matter rather than form distinguishes it inside that family. Public descriptions emphasize concerns particular to AI: impacts of AI systems on people and on society and not only risks to the organization, the AI lifecycle including data and third-party components, and documentation of how AI is used and by whom.

Two design choices in it are strengths that get mistaken for gaps. The first is neutrality toward technology. As publicly described, it does not prescribe model architectures, runtime mechanisms, thresholds, or metrics; it asks that an organization determine what to monitor and measure and demonstrate that it did so. The breadth is deliberate.

Scope is the second choice. The standard is organizational, governing how an organization manages AI rather than how a given model behaves in a given interaction. Certification accordingly says something real about the management system behind a deployment, a different claim from any statement about a particular action by a particular agent. Neither choice is a defect, and nothing below is offered as a criticism of either.

Computing the Departure Rather Than Logging It

The filed chapter treats a departure from a declared constraint as a quantity rather than a binary fault, and states why: a mechanism in which every departure is a fault carries no quantity for the need or urgency of the action attempted, and so has no input from which a permission condition could be computed. What it supplies instead is a quotient. Deviation likelihood is the difference between a need quantity and a dynamic ethical threshold, divided by the product of an empathy weighting and a self-esteem aggregate.

The need quantity is a quantified semantic urgency, computed and not declared, carried per integrity scope and per categorical type: resource scarcity, affective overload, identity dissonance, relational deficit. It accumulates by a declared base rate under a product of four declared modulators and is written to zero upon a recorded resolution of that type. Elapsed time alone neither raises nor lowers it.

Resolved per scope when a proposed mutation is evaluated, the dynamic ethical threshold is the minimum condition for deviation: the greater of a floor and the sum of a base threshold from the policy reference field, a context-sensitive adjustment scaled by the severity of the recorded context, and a historical adjustment computed from the count of permitted deviation records within a declared window. Each adjustment is signed and bounded, and the floor is not reachable from below.

Two resistances occupy the denominator. The empathy weighting aggregates the anticipated semantic impact of the mutation across personal, interpersonal, and global scopes, each scope quantity rising as impact rises against the tolerance declared for that scope. The self-esteem aggregate tracks coherence between declared intents and executed actions, its decrement on dissonant execution being entropy-weighted.

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 and no other mutation the policy object forbids. The signed policy object is not nullified and not amended and remains authoritative; the conduct deviates from it through a recorded override.

Admission appends a permitted deviation record: a deviation trigger signature for the need-threshold imbalance, a context hash of the values then in force, an integrity displacement vector quantifying degree and direction of deviation across the three integrity components, an identification of the policy constraint overridden, and a

restoration status. A party reading the lineage field reconstructs that the mutation was admitted under the permission condition, and by what quantities, rather than meeting an unexplained fault. Such a record increments no refusal counter and writes no authorization gate to a withheld state.

Restraint is written down as well. Where the prerequisites hold and the agent nonetheless does not deviate, the non-deviation is a suppression, and a dissonance buffer entry receives a violation signature naming the constraint implicated, an affective context vector, a suppression flag, and a divergence score. That score is the quotient as computed at the moment of suppression, stored with its four operands and not recomputed.

Above all of it sits a budget: a declared deviation deductible, a declared aggregate retention, and a retention register the agent maintains. The entropy-weighted harm coefficient of each permitted deviation is drawn first against the deductible, without any determination of whether the deviation was well founded. Harm exceeding the deductible creates a reparation arc, which reverses nothing; whether that arc can be discharged, and by whom, depends on where the deviation landed. The undischarged amount accumulates in the register, and where it exceeds the aggregate retention the permission condition is foreclosed: a deviation likelihood exceeding unity thereafter produces the withholding outcome and not the admission outcome, until discharge returns the register below the retention.

Organizational Layer, Runtime Layer

Both put a declared policy at the center. The standard asks an organization to set one and show conformance to it; the filed architecture takes a signed policy object as what conduct is measured against and what a permitted deviation record names as overridden. Convergence at that categorical level is why the two fit together rather than compete.

Divergence is a matter of layer. A management system standard, as publicly described, states what an organization must demonstrate and leaves open how. The filed architecture is a how: it runs inside a single agent's cognitive cycle, evaluates each proposed mutation on its own prerequisites, and appends a record per event. Nothing in it establishes a management system, and the neutrality just described calls for no particular runtime.

Three consequences follow, each a difference of role rather than of quality.

Cadence. Assessment against a management system standard is, as publicly described, periodic and performed by people. The quotient is evaluated continuously as part of the agent's cognitive cycle. Where a proposed mutation implicates more than one scope, one quotient is computed for each, and the unity condition is satisfied only where satisfied for every implicated scope.

Unit of record. A nonconformity in a management system is an organizational finding written by a person. A permitted deviation record is an appended entry naming the constraint overridden and carrying the quantities behind the admission, beside an entry record identifying the mutation, the implicated scopes, and the deviation likelihood at entry, and a termination record identifying which of three conditions ended the state and whether the mutation was admitted. Elapsed time terminates no such state.

Direction of the remedy. Here the cross-organization case separates most sharply. Corrective action inside a management system is directed at the organization's own process. In the filed architecture, an arc whose deviation is recorded against the personal integrity component is self-directed and is discharged by a restorative mutation and by nothing else. Where the affected-party class resolves to an identified counterparty holding a counterparty identity record, the arc is other-directed, and a restorative mutation performed by the agent alone discharges it not at all. Where it resolves to no identified counterparty, the arc is unaddressed, undischargeable, and

accumulated in the retention register at a declared multiple. An organization can close every internal corrective action it has opened and still carry an undischarged other-directed arc.

Read as complements, the two line up. Monitoring, measurement, internal audit, and corrective action need inputs, and a technology-neutral standard leaves their production to the organization. Permitted deviation records, suppression entries with their divergence scores, the retention register, the integrity compliance score, and the mutation controller's appended response, score, threshold, and step index are candidates. None is a certification, and no runtime issues one.

Running Both Together, and What Stays Unsolved

The join between the two is the signed policy object. For a runtime record to bear on an organization's conformance case, the constraints declared under the management system have to be the ones expressed in that object, which is what a permitted deviation record names as overridden. Getting the mapping right is a governance exercise rather than an integration one, because every parameter in the chapter is declared: base rates and need ceilings, modulator mappings, scope and impact coefficients, tolerances, adjustment bounds, the deductible, and the aggregate retention. The architecture computes; it does not choose these values. Declared carelessly, they yield a precisely computed number resting on poor inputs.

One declaration belongs on the governance function's desk. The direction of the historical adjustment is itself declared, and both directions are disclosed: in one embodiment the threshold rises as prior permitted deviations accumulate, and in a further embodiment the declared direction lowers it, reflecting normalized deviation patterns. That is a control decision, and it sits with whoever owns the policy.

What the architecture leaves undone deserves equal weight.

It certifies nothing and does not shrink what certification asks for. Leadership commitment, competence and awareness, supplier and third-party arrangements, documented procedures, internal audit, and management review are, as publicly described, organizational work no runtime performs.

Merit goes unadjudicated. The deductible is drawn without any determination of whether the deviation was well founded, and the chapter settles that question nowhere else. The determination sits outside the mechanism, and whoever the organization assigns it to receives operands rather than a paragraph.

An other-directed arc stays open. Where the affected party is an identified counterparty, discharge is governed by the counterparty-facing chapter of the same application, and the agent acting alone accomplishes nothing toward it. Cross-organization deployments should plan around that case first. They should also size retention capacity in advance, since every increment, resolution write, state entry, termination, and controller response is appended with the event that occasioned it.

Disclosure Scope

The architecture described here is disclosed in U.S. Provisional Application No. 64/117,812, principally at the chapter on integrity quantities and permitted deviation.

Disclosed there: the scoped integrity vector and its coupling functions; the need quantity with its categorical types and accumulation rule; the dynamic ethical threshold with its floor and bounded adjustments; the self-esteem aggregate and the entropy-weighted harm coefficient; the empathy weighting resolved per scope; the deviation likelihood quotient and the unity condition; the deviation-preparation state with its entry and termination records; the permitted deviation record; the dissonance buffer entry recording suppression; reparation arcs, the integrity compliance score, and the mutation controller's ordered response set; the restorative mutation; and the deviation deductible, the retention register, and the aggregate retention.

Disclaimed: any numeric value for any declared parameter, any implementation language, storage engine, or deployment topology, any conformity assessment or certification determination, and the description of standards practice above, offered as general professional background rather than as disclosure. The application is pending, and nothing here asserts that any party infringes or requires a license.

References to ISO/IEC 42001 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)

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)