Federated Decision Exchange

Sharing What Agents Get Wrong, Without Sharing What They Touched

A companion paper to Decision Accounting. Where that paper sets out how a single enterprise accounts for the decisions it delegates to AI agents, this one addresses what becomes possible when those records cross firm boundaries — and why every previous attempt to share operational failure data has succeeded or failed on three conditions rather than on goodwill. This is an early working draft, published to invite disagreement. The mechanism is sketched, not specified.

Dev Gandhi · Actee · September 2026 · v0.9, working draft

1. The Asymmetry Worth Fixing

An institution learns about agent failure the way it learns about most operational failure: by suffering it. The first firm to discover that a particular agent pattern — a prompt-template change that induces runaway tool-calling against financial reporting systems, a retrieval configuration that silently drops the most recent quarter, a delegation ceiling that accumulates rather than caps — pays the full cost of the discovery. Every other institution running a comparable configuration pays it again, separately, later.

This is a solved problem in adjacent domains. Financial crime, cyber defence and aviation safety all converged on the same insight: incidents observed by one institution predict incidents at the next, and sectors that share signatures outperform sectors that hoard them. The information-sharing bodies built on that insight — sectoral ISACs, confidential aviation safety reporting — are among the most successful voluntary institutions in modern industry.

Agent operations have no equivalent, and the reason is not reluctance. It is that the precondition has been missing.

2. Three Conditions, Not Goodwill

Institutions do not share operational failure data because it would be collectively rational. They share when three conditions hold simultaneously. Each has failed, historically, for AI operations.

- **A record worth sharing exists —** Financial crime typologies, cyber indicators of compromise and aviation occurrence reports are all standardized artifacts. Before the artifact existed, sharing meant prose, and prose does not aggregate. AI operations have had no such record — which is precisely what a Decision Ledger supplies as a byproduct rather than as an additional obligation. An institution keeping decision accounts already holds the raw material; the exchange asks it to export a derivative, not to instrument anything new.
- **Sharing carries no competitive or liability cost —** This is the binding constraint. An institution will not disclose that its agents mispriced a book, mishandled a client, or breached a delegation — not because it fails to see the collective benefit, but because the disclosure is attributable and the liability is real. The answer cannot be a promise of discretion. It has to be structural: the shared artifact must carry the failure mode and not the firm, the client, or the data.
- **A credible convening body owns the standard —** Voluntary standards in wholesale markets do not emerge from vendors, and they rarely emerge from regulation. They emerge from member-funded, consensus-based industry bodies with the standing to convene competitors and the machinery to take a proposal from taskforce recommendation to published standard with committed member adoption. That machinery exists, has recent precedent in post-trade operational standards, and is the natural home for a Decision Event schema and a mandate taxonomy. What it currently lacks is a proposal.

3. What Is Actually Shared

The unit of exchange is a decision-incident signature: a structural description of a failure pattern, derived from ledger records, with identity, entity references and payload content removed at derivation rather than redacted afterwards.

A signature answers what happened in the shape of the failure, not in the particulars of the firm. Illustratively:

> "An agent class holding a pricing mandate, operating through a vendor platform of type X, exhibited progressive threshold drift following a prompt-template revision. Individual actions remained within per-action limits throughout. Cumulative effect on a margin-class metric exceeded the aggregate ceiling on day nine. Detected post-action; no pre-action gate available on this runtime."

Note what that record contains and what it does not. It contains the mandate class, the runtime characteristic, the failure mechanism, the detection mode and the time-to-detection — everything another institution needs to check whether it is exposed. It contains no firm, no client, no product, no data, and no figure that could be triangulated back to a balance sheet.

The cryptographic design of Principle VII of the parent framework exists for this reason: to make a contribution verifiable without being attributable. A member can prove that a signature was derived from a genuine ledger record held by a genuine member, without revealing which member. Aggregate quantities carry privacy-preserving treatment so that a sufficiently determined recipient cannot reconstruct a contributor from a sequence of contributions.

4. What a Member Gets Back

- **Early warning —** A member receives notice of failure modes observed elsewhere in its sector, matched against its own fleet configuration — the agent classes it runs, the mandate types it has issued, the runtimes it depends on. The value is asymmetric in the member's favour: it learns from every other member's incidents and contributes its own.
- **Sector benchmark —** An institution can see whether its detection rate, its time-to-detection, and its breach frequency are typical for its adoption stage and agent mix. Today no such denominator exists, and every institution privately suspects it is the only one with the problem.
- **Systemic visibility —** A pool of signatures across many institutions makes visible what no single institution can see: that a particular vendor runtime lacks a pre-action interception boundary, or that a class of mandate is systematically written in a way that permits accumulation. This is the argument for the exchange that is hardest to make internally and most valuable externally.
- **Concentration visibility —** An institution can measure how much of its own agent workforce depends on one model, one vendor or one runtime. It cannot discover that its entire sector has converged on the same three. Concentration is invisible from inside a single firm by construction, and it is the exposure most likely to produce correlated failure across an industry on the same day. A signature pool carrying runtime and model-class attributes — without carrying firms — is the only mechanism we can identify that makes sector-level concentration measurable at all.
- **Supervisory and underwriting standing —** Membership of a functioning exchange, with contribution and consumption evidenced, is itself an artifact a supervisor or an insurer will recognise — the same way participation in sectoral cyber information sharing has become a de facto expectation.

5. Sequencing — Stated Plainly, Because the Reverse Order Fails

No firm will implement a decision ledger in order to share. The internal case must stand entirely on its own: board answerability, examination readiness, incident attribution, and learning which delegations are working. Federation is what becomes available once institutions happen to keep compatible books.

The order therefore runs: institutions adopt the discipline for internal reasons; a convening body adopts the Decision Event schema and mandate taxonomy as a published standard; early members contribute signatures bilaterally or in a small pool; the pool reaches the density at which matching becomes useful; participation becomes an expectation. Each step is worth taking on its own terms, which is the only property that makes a voluntary standard survive.

We state this because the temptation in a paper of this kind is to describe the endpoint and imply the path. The endpoint is not the argument. The first step is.

6. Open Questions

This draft is published to be argued with. The questions we cannot yet answer:

1. Which body should own the schema — an existing sectoral standards organization, a cross-sector coalition, or a new entity? Each has precedent and each has failure modes.
2. How thin can a signature be and remain useful? Too rich and it becomes attributable; too sparse and it matches nothing.
3. What is the minimum pool density at which matching produces value, and how does an exchange survive the period before it reaches that density?
4. Should contribution be a condition of consumption? Reciprocity solves free-riding and suppresses participation; the trade-off has been resolved differently in different sectors.
5. How does an exchange handle a signature that implicates a named vendor runtime? The information is the most valuable in the pool and the most legally fraught — and it is also what concentration measurement requires.

We would rather publish these unresolved than resolve them alone. Comments to info@actee.ai.

This paper is a companion to Decision Accounting: A Framework for AI-Agent Decision Accountability in the Enterprise, which sets out the discipline this exchange depends on.