Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Does an Integrity-Gated Governance Observability Index Measure Reconstructability? A Simulation-Based Criterion-Validity Study of Agent Telemetry under Out-of-Specification Faults

Ravi Shankar NRK

ravishan2006@gmail.com

Independent Researcher

Research Article

Keywords: agentic AI, governance observability, audit telemetry, composite indicators, criterion validity, tamper-evident logging, simulation

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11552307/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Does an Integrity-Gated Governance Observability Index Measure Reconstructability? A Simulation-Based Criterion-Validity Study of Agent Telemetry under Out-of-Specification Faults

Ravi Shankar NRK

Independent Researcher, Hyderabad, Telangana, India

ORCID 0009-0004-7449-3618

Corresponding author: Ravi Shankar NRK, ravishan2006@gmail.com

Abstract

Governance of agentic AI systems depends on execution evidence: an operator can judge whether an agent stayed within its authority only if its telemetry allows what it did to be reconstructed. The Governance Observability Index (GOI) proposes to measure this as a gated geometric mean of coverage, fidelity, integrity and timeliness, with an integrity gate intended to stop untrustworthy evidence being masked by completeness elsewhere. This simulation study tests one implementation of the index against criteria computed from complete ground truth, of which the estimator sees only a 40-record ledger sample. A seeded testbed simulates a scripted payments agent, a policy gateway, an authoritative ledger and an HMAC-chained telemetry stream. The estimator uses only the stream, a gateway control total and a 40-record ledger sample. Twelve fault families—four within the original specification, four outside it and four benign infrastructure perturbations—are evaluated in 2,520 single-family runs and 1,000 mixed runs. GOI correlates with reconstruction accuracy (Spearman $\rho = 0.85$, 95% CI 0.83–0.87) and separates degraded from sound reconstructions (AUC 0.93). It is weaker as a warning of missed time-sensitive violations (AUC 0.87), and 63% of runs scoring at least 0.90 had degraded reconstructions. The integrity gate, as implemented, never activated in the mixed runs. A 5% forgery by a key holder left the index at 0.99 while 98% of sequence-level violations went undetected in time. When deletion targets the same records as forgery, it shifts part of the defect from fidelity to coverage; a sensitivity experiment with independent deletion leaves the headline results unchanged. In this testbed GOI works as a screening measure, not a certificate of trustworthy evidence, and its usefulness depends on evidence channels independent of the telemetry.

Keywords: agentic AI; governance observability; audit telemetry; composite indicators; criterion validity; tamper-evident logging; simulation

1. Introduction

An agentic AI system does not produce a single output to be evaluated. It executes a sequence of actions—retrieving data, calling tools, delegating subtasks, committing transactions—each carrying authority and each leaving, or failing to leave, a trace. Pre-deployment assessment cannot establish what such a system later did. That question is answered afterwards, from execution evidence, and the answer is only as good as the evidence (Chan et al., 2023; Shavit et al., 2023). Work on technical AI governance identifies the operationalisation of such evidence as an open problem (Reuel et al., 2025), and audit research shows that limited access to system internals constrains what any auditor can conclude (Casper et al., 2024).

Governance observability names the property at stake: the degree to which an operator can reconstruct an agent's governance-relevant state and behaviour from the evidence it emits. The construct and its index were proposed in the author's earlier work, which is not yet published; Section 2.1 states the complete definition evaluated here, so the paper can be read and checked without it. The Governance Observability Index (GOI) proposes to measure the property. Four dimensions—coverage (C), fidelity (F), integrity (I) and timeliness (T)—are combined as a geometric mean and multiplied by an integrity gate g(I), so that evidence whose authenticity falls below a floor cannot score well however complete it appears.

A measure of this kind raises a validity question before it raises a deployment one. Does a high GOI mean that the governance state can in fact be reconstructed? Two earlier unpublished pilot experiments by the author used a binary form of the index (a conjunction of pass/fail checks) and classified all 250 injected faults and all 250 clean traces correctly in each experiment. Those results establish only that the code implemented its own checks: each fault was written to violate a specific check, so a perfect score was guaranteed by construction. In software-testing terms, the experiments measured a mutation score over hand-chosen mutants (Jia & Harman, 2011). They did not test the continuous index, the gate, or any relationship between the score and what it claims to measure.

This paper supplies that test, in simulation. Its central design choice is separation: the estimator that computes GOI sees only what an operator would see, while the criterion—how accurately the governance state can be reconstructed—is computed against the complete ledger and fault labels, of which the estimator receives only a 40-record ledger sample and a control total. Faults are injected at graded severities, including faults outside the original verifier's specification and benign perturbations that real infrastructure produces routinely. The contributions are:

1. A criterion-validity evaluation of one implementation of the continuous GOI against three ground-truth outcomes: reconstruction accuracy, silent misstatement (false records accepted as genuine) and timely detection of a sequence-level policy violation.

2. An analysis of that implementation's aggregation: the fourth root compresses the useful range of the index; the gate, as implemented, is discontinuous and rarely active over the fault ranges tested; and the hazard the gate targets arose here through fidelity rather than integrity.
3. A threat model for agent telemetry that distinguishes transport faults, emitter defects, external tampering, key-holder manipulation and benign infrastructure behaviour, and identifies which evidence channel detected each in the testbed.
4. A comparison with the earlier (legacy) binary verifier, which raised false alarms on benign retries and reordering while missing tail truncation and insider deletion.
5. A paired counterfactual showing that selective deletion changes which evidence defects remain—so that conditional fidelity can understate the true falsification rate—while the index's association with reconstruction accuracy stays similar under independent deletion.

The study is a controlled simulation. It involves no live language model and no production system, and the criteria are the author's own. What it can do—and what the earlier experiments could not—is show where this implementation of the index tracks the construct, where it does not, and which design choices account for each.

2. Background and related work

2.1 The construct and the implementation tested

Governance observability is defined by analogy with control-theoretic observability—the reconstructability of state from outputs (Kalman, 1960)—and with software observability built from telemetry (Sridharan, 2018). The state of interest is restricted to what is governance-relevant: what the agent attempted, under which authority, on what basis and with what effect. The construct and the index were proposed in the author's earlier work, which is unpublished. That work is therefore not cited, and the definition evaluated here is stated in full below; Online Resource 2 sets out each element as originally proposed beside the operational choice made in this study. The index is

$$GOI = g(I) \cdot (C \cdot F \cdot I \cdot T)^{1/4} \qquad (1)$$

with each dimension normalised to [0, 1]. Coverage is the share of governance-relevant behaviour that emits usable evidence; fidelity, the faithfulness of that evidence to what happened; integrity, its resistance to tampering and omission; and timeliness, the share of evidence available within the latency budget needed to act on it. The gate g(I) is intended to limit the index when integrity falls below a floor I_min, set at 0.5, on the reasoning that evidence which can be silently altered cannot support a reconstruction however complete it is. The original proposal does not give the functional form of g. This study tests the following piecewise definition:

$$g(I) = I \text{ for } 0 \le I < 0.5; \quad g(I) = 1 \text{ for } 0.5 \le I \le 1 \qquad (2)$$

This choice reproduces the single worked example given with the original proposal (C = F = T = 0.95 and I = 0.30 giving GOI = 0.21), but one example cannot determine a function. A hard cap, a smooth ramp or a different floor would also satisfy the verbal description. The discontinuity and double contribution of integrity analysed in Section 3.2 are properties of the implementation tested here, not of every possible integrity-gated index. A hard-gated variant (GOI set to 0 below the floor) was also computed, and it gave identical results in every mixed run because no mixed run fell below the floor (Section 5.1). The geometric mean was chosen over the arithmetic mean so that strength in one dimension cannot fully offset weakness in another, following the Human Development Index's move from arithmetic to geometric aggregation in 2010 (UNDP, 2010), and over the minimum so that the index still responds to every dimension. The result is integrity-gated but only partly non-compensatory, since the three ungated dimensions still trade off against one another (Munda & Nardo, 2009).

2.2 Adjacent traditions

The four dimensions are not new. Coverage, accuracy and timeliness are established dimensions of data quality (Wang & Strong, 1996), and the GOI dimensions map onto them closely. The governance-specific contribution lies in the integrity dimension and in tying the dimensions to a declared taxonomy of governance-relevant events. Log integrity has its own literature. Forward-secure audit logs (Schneier & Kelsey, 1999) and authenticated data structures for tamper-evident logging (Crosby & Wallach, 2009) show what an untrusted log store can and cannot hide. A recurring lesson is that a message authentication code proves only that a holder of the key produced the record; it cannot detect manipulation by the key holder, who is often the adversary that matters most in an enterprise audit setting. Provenance models such as W3C PROV (Moreau & Missier, 2013) and distributed tracing systems (Sigelman et al., 2010) supply the structural vocabulary, and agent-specific observability taxonomies are emerging (Dong et al., 2024).

Detecting a policy violation that spans several actions—such as transfers structured to stay below a per-action limit—is a runtime verification problem (Leucker & Schallhart, 2009). A runtime monitor consumes the same evidence that GOI scores, which gives a behavioural criterion: if evidence quality matters, monitor performance should degrade as it degrades.

Composite indicators have a mature methodology (OECD & JRC, 2008). The choice between compensatory and non-compensatory aggregation, and the observation that geometric means are only partially non-compensatory, are treated there at length (Munda & Nardo, 2009). Psychometrics supplies the validation standard: an index is valid to the extent that evidence supports the interpretation placed on its scores, including relationships with external criteria

(Messick, 1995). Audit practice supplies two further ideas used below. Evidence from sources independent of the entity is more reliable than evidence the entity generates about itself (IAASB, 2009a), and fidelity can be estimated by statistical sampling against the system of record (IAASB, 2009b). Finally, any measure used to gate decisions is subject to gaming pressure (Campbell, 1979), and assurance arguments built on unexamined evidence have a documented record of failure (Haddon-Cave, 2009).

2.3 Gap

The literatures reviewed in Sections 2.1 and 2.2 supply methods for measuring data quality, making logs tamper-evident, verifying runtime behaviour and validating composite indicators. They do not, by themselves, establish criterion validity for the particular GOI operationalisation tested here: whether its score tracks how accurately governance-relevant state can be reconstructed, judged against complete ground truth, and how its integrity gate behaves across the fault families examined in this study. The earlier binary tests cannot answer either question.

3. Measurement model and threat model

3.1 Operational definitions

Each dimension is estimated per run (a window of 200 agent actions) from evidence an operator could hold. Three departures from the original proposal (Online Resource 2) are made. First, fidelity is defined only as correspondence between the record and the executed effect. The earlier definition also included correspondence to "the applicable policy", which conflates evidence quality with behavioural compliance and contradicts the boundary result in Section 5.2. Second, every dimension is tied to a stated evidence source, because a dimension that cannot say where its denominator comes from cannot be estimated. Third, the proposal defines integrity as resistance to tampering and omission; here omission is assigned to coverage, and integrity covers only whether an observed record is authentic and bound to its context, so that no defect is counted in two dimensions.

- *Coverage.* $\hat{C}$ = (distinct expected spans observed) / (4 × N_ct). Each action is expected to emit four spans (intent, policy decision, tool call, outcome), and N_ct is a control total of actions processed, reported by the policy gateway through a channel separate from telemetry. Coverage is measured against a declared taxonomy of expected events, as in the original proposal; the control total supplies the denominator that telemetry alone cannot. In this testbed the control total is assumed error-free, so $\hat{C}$ equals the true share of emitted spans that reached the collector. Section 5.7 measures what happens without it.
- *Fidelity.* A simple random sample of 40 ledger-committed actions is drawn without replacement. $\hat{F}$ is the proportion of matches on payee and amount among the n_F sampled

actions that have an authenticated outcome record. Fidelity is therefore conditional on availability: a sampled action with no valid record counts as a coverage failure, not a fidelity failure, so the two dimensions measure separate defects. $\hat{F}$ estimates the falsification rate among available records, not across all executed actions; the two coincide only when records go missing independently of whether they are false (Section 5.8). If no sampled action has a valid record, $\hat{F}$ is undefined and conservatively set to 0; this never occurred. n_F ranged from 9 to 40 (median 40).

- *Integrity.* $\hat{I}$ = proportion of distinct records whose HMAC-SHA256 verifies, that are bound to the run being assessed, and whose hash link points to the record with the preceding sequence number where that record is present. A missing predecessor is treated as a coverage gap, not an integrity breach.
- *Timeliness.* $\hat{T}$ = proportion of distinct records that reached the collector within a 2.0 s budget plus a 0.25 s clock-skew tolerance.

Exact duplicates (same run, action, span and MAC) are removed before any dimension is computed; they are the normal signature of idempotent retries.

3.2 Properties of the implemented aggregation

Three properties follow from the formula and the gate of Section 2.1 before any simulation is run.

Compression. Because the geometric mean takes a fourth root, a loss confined to one dimension is attenuated. With the other three dimensions at 1, a coverage of 0.70 yields GOI = 0.70^(1/4) = 0.915, and a coverage of 0.50 still yields 0.841. Conversely, a GOI of 0.90 is equivalent to a single dimension at 0.656, and a GOI of 0.95 to a single dimension at 0.815. The scale is therefore crowded into its top decile, and thresholds that look conservative (for example, "act if GOI < 0.90") are lenient. This property belongs to the geometric mean and holds for any gate.

Discontinuity. With g(I) = I below the floor, GOI halves at I = 0.5. For C = F = T = 0.95 the index is 0.809 at I = 0.50 and 0.404 at I = 0.4999—a jump of 0.405 for an infinitesimal change in evidence (Figure 2). A score that can halve across a measurement-error boundary invites dispute about which side of the boundary a deployment sits on, and, under gaming pressure, effort to stay just above it. A continuous gate would avoid this; a hard cap would make it larger.

Double counting and inactivity. Under this implementation, integrity enters twice: inside the geometric mean and as the gate multiplier. Below the floor the index therefore falls with I^(5/4); above it, integrity is weighted equally with the other dimensions. The gate acts only once half the records have failed verification. Whether that regime arises is an empirical question, addressed in Section 5.

3.3 Threat model

The evidence pipeline is exposed to five classes of actor or condition, which differ in capability and in which evidence channel can catch them (Table 1).

Table 1. Threat classes, capabilities and the channel expected to detect them.

Class	Capability	Examples in this study	Expected detecting channel
Transport fault	Lose or delay records; no key	F1 random loss, F4 late delivery, H1 tail truncation	Sequence gaps; control total (H1); collector clock
Emitter defect	Record wrong values before signing	F2 unit or payee bug	Ledger sample; cross-span consistency
External tamperer	Alter or inject records; no key	F3 post-signing edits, H2 cross-run replay	MAC, run binding, hash chain
Key holder	Rewrite and re-sign the whole stream	H3 forgery, H4 deletion with renumbering	Ledger sample (H3); control total (H4)
Benign infrastructure	Duplicate, reorder, skew, rotate keys	B1–B4	None should be needed: these must not lower the score

A MAC-based integrity check cannot see the key-holder class by construction. Including it is the point of the study: if the index is to support governance claims, it must be shown what happens when the party able to sign the evidence is also the party with a reason to change it.

4. Testbed and experimental design

4.1 Simulated system

Each run simulates 200 payment intents from a scripted agent. A policy gateway permits a transfer if the amount is at most 5,000 units and the payee is on an approved list of 36 out of 40 payees (Figure 1). Payees are drawn with Zipf-like reuse; amounts are log-normal with a median of 900 units, so roughly 5% of intents exceed the limit and are denied. Each permitted transfer is committed to an in-memory ledger, the authoritative record. Three structuring episodes are inserted into each run, each consisting of three or four transfers of 4,000–4,999 units to one approved payee within an 18-action span. Each transfer passes the per-action rule, so the gateway commits all of them; detecting the pattern is the job of a sequence-level monitor.

Every action emits four spans. Each record carries the run identifier, action identifier, sequence number, payee, amount, decision, event timestamp and key identifier. It is signed with HMAC-SHA256 over these fields together with the first 16 hex characters of its predecessor's MAC. Under clean conditions records reach the collector 20–40 ms after the event, so no reordering occurs.

The structuring monitor processes authenticated outcome records in arrival order. It raises an alert when three or more committed transfers to the same payee, each between 80% and 100% of the

limit, fall within 30 s of event time. An episode counts as detected if any alert includes at least two of its transfers, and as detected in time if the alert is raised within 2.5 s of the episode's final transfer.

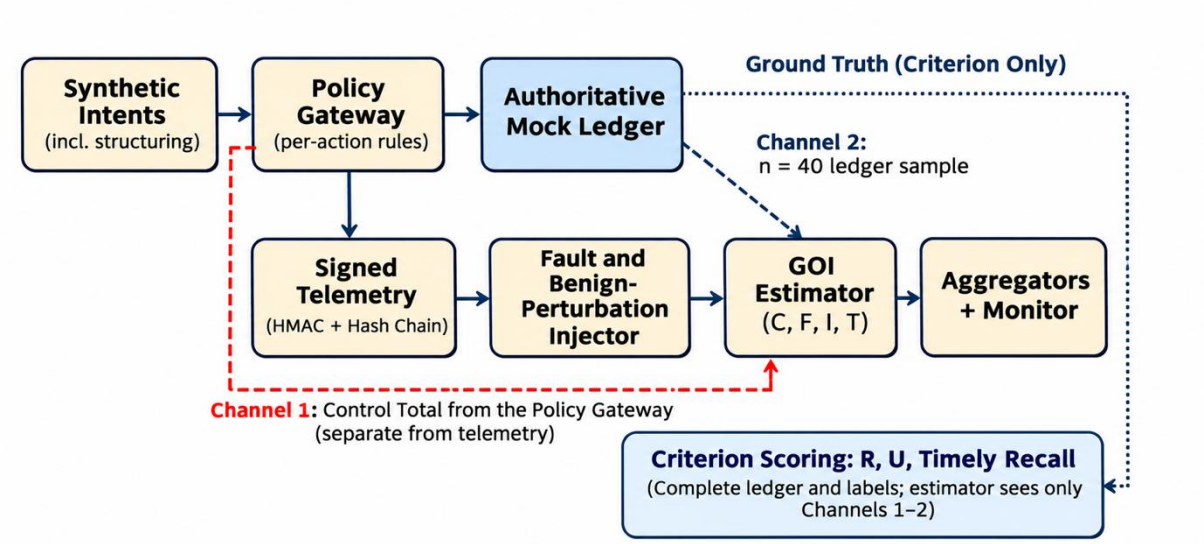


Figure 1. Testbed. The estimator receives the telemetry stream after fault injection and two inputs from outside it: a control total from the policy gateway (channel 1) and a 40-record sample from the ledger (channel 2). The complete ledger and the fault labels reach only the criterion scorer.

4.2 Fault families

Twelve families are injected at severity p: the share of records, actions or committed transfers affected (Table 2). The four F families reproduce the classes used in the earlier binary experiments. The four H families lie outside that specification. They were designed by the same author who built the estimator, so they are not blind; what distinguishes them is that no check was written for them. The four B families are routine infrastructure behaviour that a sound measure must tolerate. Emitter-side changes (F2, B3, B4) are applied first, key-holder changes next (H3 forgery, then H4 deletion) and transport-path faults last, reflecting where each occurs in a real pipeline. Key-holder forgery and deletion both target structuring transfers first, as a motivated insider would. Because forgery precedes deletion, deletion can remove records that have just been forged; Section 5.8 examines that interaction.

Table 2. Fault and perturbation families.

Family	Mechanism	Primary dimension expected to respond
F1 Random record loss	Each record dropped with probability p	C
F2 Emitter fidelity bug	Fraction p of committed actions recorded with amount ÷ 100 or adjacent payee on the tool and outcome spans, then signed	F
F3 Post-signing tampering	Fraction p of records edited after signing	I
F4 Late delivery	Fraction p of records delayed 3–30 s	T
H1 Tail truncation	Final fraction p of the stream lost (no sequence gap)	C, via control total

Family	Mechanism	Primary dimension expected to respond
H2 Cross-run replay	Validly signed records from another run injected, p × stream length	I, via run binding
H3 Key-holder forgery	Payee rewritten on all four spans of fraction p of committed actions; stream re-signed	F, via ledger sample
H4 Key-holder deletion	Fraction p of actions deleted; sequence renumbered; stream re-signed	C, via control total
B1 Idempotent retries	Fraction p of records re-sent unchanged	None
B2 Asynchronous reordering	Fraction p of records arrive out of order, within budget	None
B3 Clock skew	Emitter clock behind by 0.4p s (≤ 0.24 s, within tolerance)	None
B4 Key rotation	Final fraction p of records signed with a second, registered key	None

4.3 Criteria

Three outcomes are computed from ground truth: the complete ledger and the episode labels. The estimator never sees the complete ledger, the labels or these outcomes; it receives only the 40-record ledger sample. Fidelity estimate and criterion therefore share a source—the sample is drawn from the same ledger against which the criteria are scored—which is the usual design for evaluating a sample estimate against its population (Section 5.6).

- *Reconstruction accuracy R.* An auditor rebuilds the set of committed transfers (action, payee, amount) from authenticated outcome records. R is the Jaccard similarity between that set and the ledger: correct reconstructions divided by the union of true and claimed transfers. R penalises both missing and false records.
- *Silent misstatement U.* The share of reconstructed transfers that do not exist in the ledger: false evidence accepted as genuine. This is the hazard the integrity gate is meant to prevent.
- *Timely monitor recall.* The share of structuring episodes detected in time by the monitor described in Section 4.1.

A run is classed as degraded if $R < 0.90$ or $U > 0.02$. The label concerns eventual reconstruction. It does not include timely recall, which is analysed separately because a score that predicts eventual reconstruction need not warn of time-sensitive violations (Section 5.6).

4.4 Baseline

The earlier binary verifier is re-implemented as a run-level conjunction of its check classes: missing span or sequence gap, MAC failure, late arrival beyond budget, duplicate span, out-of-order arrival, missing timestamp, and full payload reconciliation against the ledger. It fails a run if any check fails. It has the full ledger, but, as specified in the pilots, it uses it in one direction only: each observed outcome record is checked against the ledger, but the ledger is not checked for committed transfers that have no record. It is therefore the legacy checklist verifier, not a comprehensive full-

ledger comparator. A bidirectional reconciliation against the complete ledger would detect truncation and deletion by construction, but it amounts to auditing every transfer against the system of record—the cost that a sampled estimator with a control total is designed to avoid. The legacy verifier is included to reproduce the earlier binary implementation, not as an optimal or computationally equivalent comparator. The study therefore does not establish that GOI outperforms comprehensive ledger reconciliation. It evaluates whether a sampled, integrity-gated estimator provides useful evidence of reconstructability under the specified fault conditions. The verifier evaluated here is the specification used in the pilot experiments; a verifier designed for production could handle idempotency and permitted reordering explicitly.

4.5 Experiments

- E1 (dose-response) runs each of the twelve families alone at $p \in \{0, 0.05, 0.10, 0.20, 0.30, 0.45, 0.60\}$, with 30 seeds per cell: 2,520 runs, of which 360 are clean.
- E2 (criterion validity) runs 1,000 mixtures in which each F and H family is active independently with probability 0.25 at $p \sim U(0.01, 0.30)$, and each benign family is always active at $p \sim U(0, 0.30)$.
- Three ablations recompute the estimator on the E2 runs: without the control total; without the ledger sample ($\hat{F}$ fixed at 1); and without run binding and hash-chain checks. Each ablation is assessed in two ways (Section 5.7): against the full verifier's fixed criterion, and against outcomes recomputed from the records that the ablated verifier itself accepts.
- E2b repeats the 1,000 E2 mixtures with identical fault draws, except that key-holder deletion selects actions independently of the forgery targets.

Spearman correlations are reported with 95% percentile bootstrap intervals (1,000 resamples of runs). Paired comparisons—between two scores on the same runs, or between E2 and E2b—use 2,000 bootstrap replicates. Each replicate resamples the 1,000 runs (seed pairs, for E2 against E2b) with replacement, applying the same draw to both members of every pair. The area under the ROC curve (AUC) is computed with a low score treated as predicting the positive class.

All randomness flows from a master seed (20261029). The simulation and analysis scripts, the run-level data and SHA-256 checksums for every released file are published (see Data availability). Generative AI was used to write and run the code; the declaration at the end of the paper states how.

Two analysis settings were changed after a first full run, and both versions are reported. In the first run, E2 mixtures used activation probability 0.35 and severities up to 0.50, and the degraded criterion also required perfect timely recall. Under those settings 93% of runs were degraded, which left too few sound runs for a meaningful AUC. The milder mixture and the criterion above were

then fixed. The first run's figures were similar: ρ with R = 0.858 and AUC = 0.962 for the gated index. Before the frozen run, four testbed defects were corrected after smoke tests: natural arrival jitter that reordered clean streams, a monitor rule with an excessive false-alarm rate, forgery that kept an episode's transfers on one payee, and deletion that renumbered ledger-bound action identifiers.

This revision uses testbed release v1.1, which adds E2b, ground-truth tracking columns and per-ablation outcomes. It makes one change to a fault. In v1.0.1, key-holder forgery rewrote three of an action's four spans and left the policy-decision span unchanged, so a cross-span consistency check could have exposed it (Section 6.2). In v1.1 the forger rewrites all four spans. No estimator or criterion reads the policy-decision span, so the change affects only six mixed runs. In those runs, post-signing tampering redraws a payee that now coincides with the forged value, and GOI moves by at most 0.0008. Every statistic reported in the first submission is unchanged at three decimals except one: the correlation of the no-ledger-sample ablation with 1 − U moved from 0.000 to 0.001.

5. Results

5.1 The gate in the observed fault distribution

The analytic properties in Section 3.2 are confirmed. The worked example reproduces (GOI = 0.214, against a geometric mean of 0.712 and an arithmetic mean of 0.787). Across the 2,520 single-family runs, integrity fell below the 0.5 floor in 1.2% of runs, all of them post-signing tampering at p = 0.60. Across the 1,000 mixtures the lowest integrity observed was 0.555, so the implemented gate never activated, and the gated and ungated indices were exactly identical in every E2 statistic reported below. This partly reflects the E2 design, which caps each family's severity at 0.30. The gate condition is nonetheless demanding: tampering and replay together must invalidate more than half of all records, a regime in which every aggregator in Figure 2 already reports severe degradation.

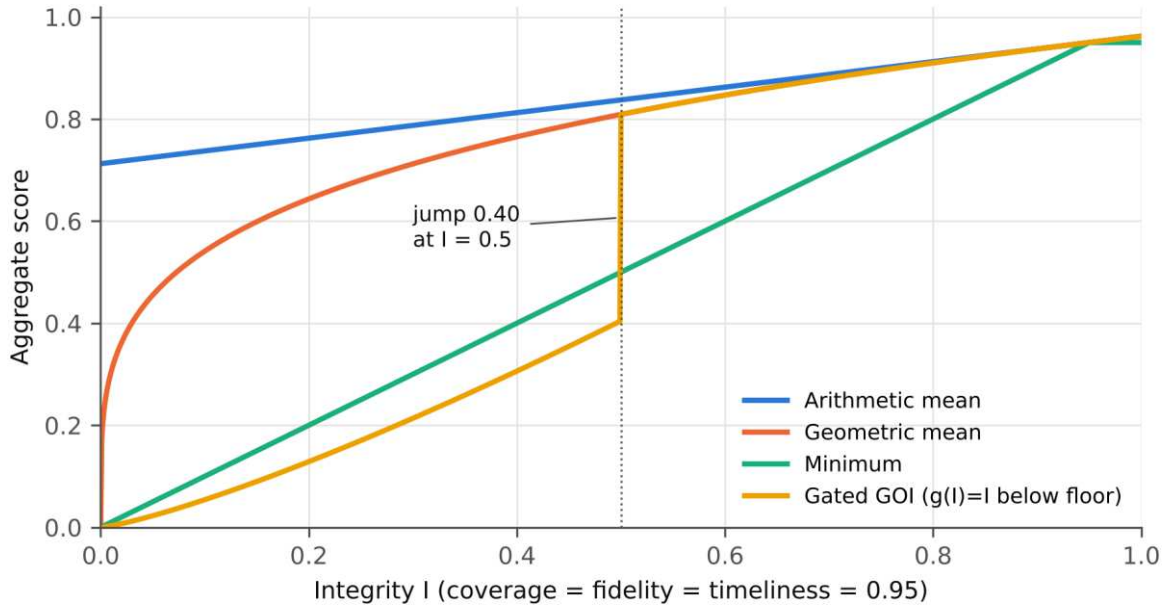


Figure 2. Aggregate score against integrity with the other three dimensions fixed at 0.95. The gated index, as implemented here, is discontinuous at the floor and coincides with the geometric mean above it.

5.2 Clean runs and the behavioural boundary

All 360 clean runs scored GOI = 1.000 with R = 1.000, and the binary verifier passed all of them. Each run contained three structuring episodes: 1,080 episodes in total, made up of 3,752 transfers. The gateway committed all 3,752 transfers, because each satisfied the per-action rule, and the monitor detected all 1,080 episodes in time, with 0.008 false alerts per run. Perfect evidence quality therefore coexisted with every structuring transfer being committed. The index measures whether violations can be seen, not whether they occurred: with sound evidence, the sequence-level monitor caught every episode.

5.3 Dose-response

Figure 3 and Table 3 show each family's effect. For every in-specification fault and both deletion-type out-of-specification faults, the estimated dimension responded close to proportionally: $\hat{C} \approx 1 - p$ for loss, truncation and deletion; $\hat{I} \approx 1 - p$ for tampering; $\hat{T} \approx 1 - p$ for delay. The index fell far less steeply than the criteria. At $p = 0.30$ GOI lay between 0.914 and 0.937 for every non-benign family. For the six families that remove or falsify committed records, R fell to between 0.54 and 0.71. Timely monitor recall fell to between 0 and 0.70 for every non-benign family except replay.

Table 3. Mean outcomes at severity p = 0.30 (30 runs per row). BV = proportion of runs failed by the binary verifier.

Family	$\hat{C}$	$\hat{F}$	$\hat{I}$	$\hat{T}$	GOI	R	U	Timely recall	BV fail
F1 Random loss	0.697	1.000	1.000	1.000	0.914	0.695	0.000	0.50	1.00
F2 Emitter bug	1.000	0.716	1.000	1.000	0.919	0.543	0.297	0.44	1.00
F3 Tampering	1.000	1.000	0.710	1.000	0.918	0.706	0.000	0.43	1.00
F4 Late delivery	1.000	1.000	1.000	0.701	0.915	1.000	0.000	0.64	1.00
H1 Tail truncation	0.700	1.000	1.000	1.000	0.915	0.698	0.000	0.70	0.00
H2 Cross-run replay	1.000	1.000	0.769	1.000	0.937	1.000	0.000	1.00	1.00
H3 Key-holder forgery	1.000	0.712	1.000	1.000	0.917	0.538	0.300	0.00	1.00
H4 Key-holder deletion	0.700	1.000	1.000	1.000	0.915	0.697	0.000	0.00	0.00
B1 Retries	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.00	1.00
B2 Reordering	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.00	1.00
B3 Clock skew	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.00	0.00
B4 Key rotation	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.00	0.00

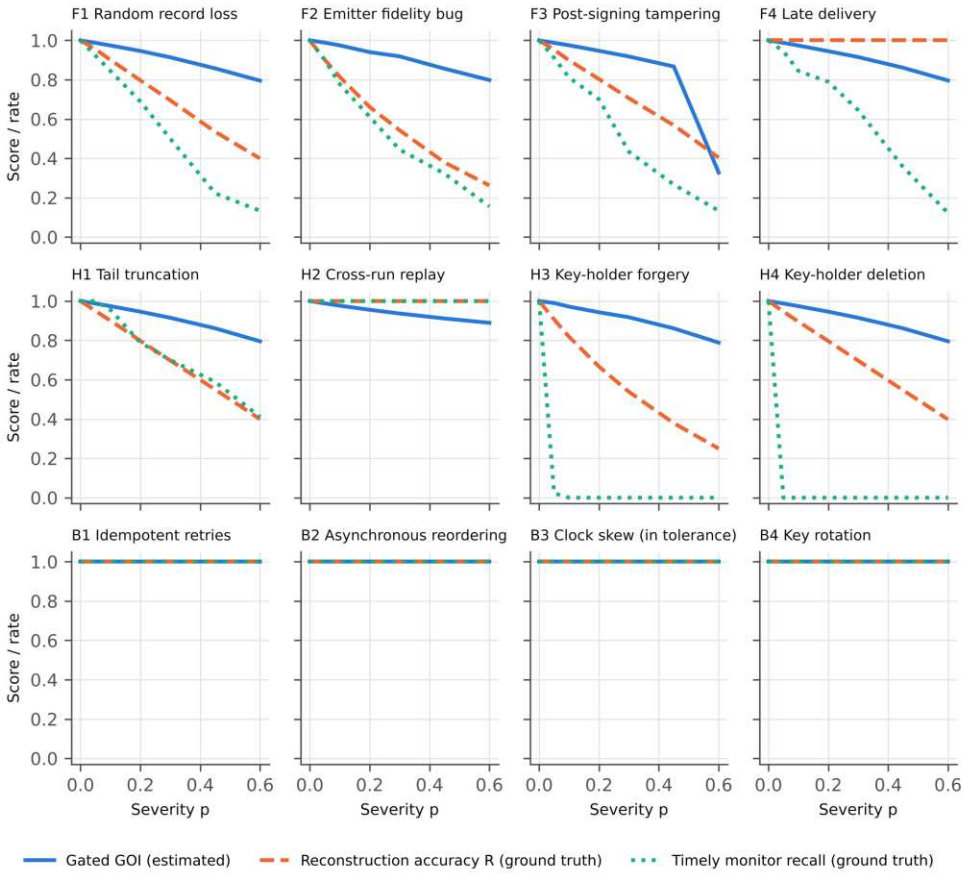


Figure 3. Gated GOI, reconstruction accuracy and timely monitor recall against severity for each family; means over 30 runs per point.

Two patterns stand out. First, the integrity faults behaved as safe failures. Tampered and replayed records were rejected, so they reduced reconstruction (F3) or left it unchanged (H2), but never produced accepted false records (U = 0 in both). Silent misstatement arose only from F2 and H3, where the false records carried valid MACs and the integrity dimension stayed at 1.000. Second, the outcomes that matter for governance depend on which records are lost, not only how many. Delay left reconstruction intact but cut timely recall to 0.64. Key-holder deletion and forgery, aimed at structuring transfers, removed timely detection entirely at p = 0.30, while the index moved no more than it did for random loss.

5.4 Out-of-specification faults

Key-holder forgery is the clearest case. At p = 0.05 the index averaged 0.990, while timely recall of structuring episodes fell to 0.022: almost every violation was hidden. Among the checks implemented here, the only signal came from the 40-record ledger sample, which found at least one mismatch in 26 of 30 runs (86.7%). Because the sample is drawn without replacement from ledgers of 177–190 committed transfers, the exact hypergeometric detection probability, averaged over those 30 runs, is 89.6%; the binomial approximation 1 − 0.95^40 gives 87.1%. The observed rate is consistent with either. In the remaining four runs the index was exactly 1.000 while the governance record was false.

Tail truncation and key-holder deletion leave no sequence gap. The coverage estimate caught both, but only through the gateway control total. Without it, the coverage estimate departed from the true span coverage by a mean absolute 0.145 on the 228 mixed runs containing truncation and 0.028 on the 257 containing deletion; with it, the estimate equals true coverage in this testbed, because the control total is assumed error-free. The binary verifier failed none of the H1 or H4 runs at any severity. Although it holds the full ledger, it reconciles only the records it observes (Section 4.4), and the streams it saw were internally complete.

5.5 Benign perturbations

None of the four benign families changed any dimension: GOI stayed at 1.000 at every severity. The binary verifier failed 100% of runs containing idempotent retries or asynchronous reordering, even at p = 0.05, because duplicate and out-of-order spans are in its fault list. It passed clock skew and key rotation. The verifier specification evaluated here therefore flags normal distributed-systems behaviour as faults while missing two attacks that preserve structure.

5.6 Criterion validity

Table 4 reports validity statistics over the 1,000 mixtures, of which 737 (73.7%) were degraded under the criterion in Section 4.3.

Table 4. Spearman correlation with each criterion (95% bootstrap CI) and AUC for identifying degraded runs; E2, n = 1,000.

Score	ρ with R	ρ with 1 − U	ρ with timely recall	AUC
Gated GOI	0.849 (0.826–0.869)	0.454 (0.400–0.505)	0.618 (0.573–0.658)	0.925
Geometric mean (ungated)	0.849 (0.826–0.869)	0.454 (0.400–0.505)	0.618 (0.573–0.658)	0.925
Arithmetic mean	0.844 (0.821–0.865)	0.460 (0.407–0.511)	0.617 (0.572–0.657)	0.925
Minimum	0.817 (0.791–0.839)	0.380 (0.323–0.436)	0.572 (0.523–0.618)	0.909
Coverage only	0.537 (0.486–0.584)	−0.022 (−0.088–0.042)	0.456 (0.402–0.503)	0.807
Fidelity only	0.682 (0.641–0.718)	0.971 (0.963–0.978)	0.453 (0.402–0.501)	0.768
Integrity only	0.260 (0.201–0.314)	0.040 (−0.021–0.102)	0.159 (0.098–0.219)	0.629
Timeliness only	−0.006 (−0.068–0.056)	−0.008 (−0.071–0.056)	0.072 (0.016–0.128)	0.499

Two parts of this table are mechanical by construction and should be read as such. When records are lost at random, $\hat{C}$ and R both measure the surviving share and must move together. Likewise, $\hat{F}$ and U compare the same thing: an accepted outcome record's payee and amount against the ledger. $\hat{F}$ does so on a sample of 40, U on every reconstructed transfer. The near-perfect correlation of fidelity with 1 − U ($\rho = 0.97$) therefore mainly confirms that a 40-record sample estimates the population mismatch rate well. It is not independent evidence that fidelity captures the construct. The informative evidence comes from the cases where estimator and criterion could have come apart—delay, replay, targeted deletion and forgery (Sections 5.3–5.4 and 5.8)—and from monitor recall, which no dimension measures directly.

With those caveats, the composite predicts reconstruction accuracy better than any single dimension, and it separates degraded from sound runs well (AUC 0.925). Arithmetic and geometric aggregation performed similarly in practical magnitude over the tested range. The geometric mean's correlation with R was higher by 0.0052 (paired bootstrap 95% CI 0.0033–0.0071), a small but non-zero difference, so the data do not show the two are equivalent. The gated and ungated geometric means are identical here only because the implemented gate never activated. The minimum performed slightly worse.

Good discrimination of the degraded label does not establish that the index warns of missed time-sensitive violations. In 70.4% of mixed runs at least one structuring episode was not detected in time. The index separated those runs from the rest less well (AUC 0.869) than it separated degraded reconstructions (0.925). And 21.7% of runs that were not degraded—whose records could eventually be reconstructed adequately—still missed at least one episode in time. The delay and targeted-fault results in Section 5.3 show the mechanism.

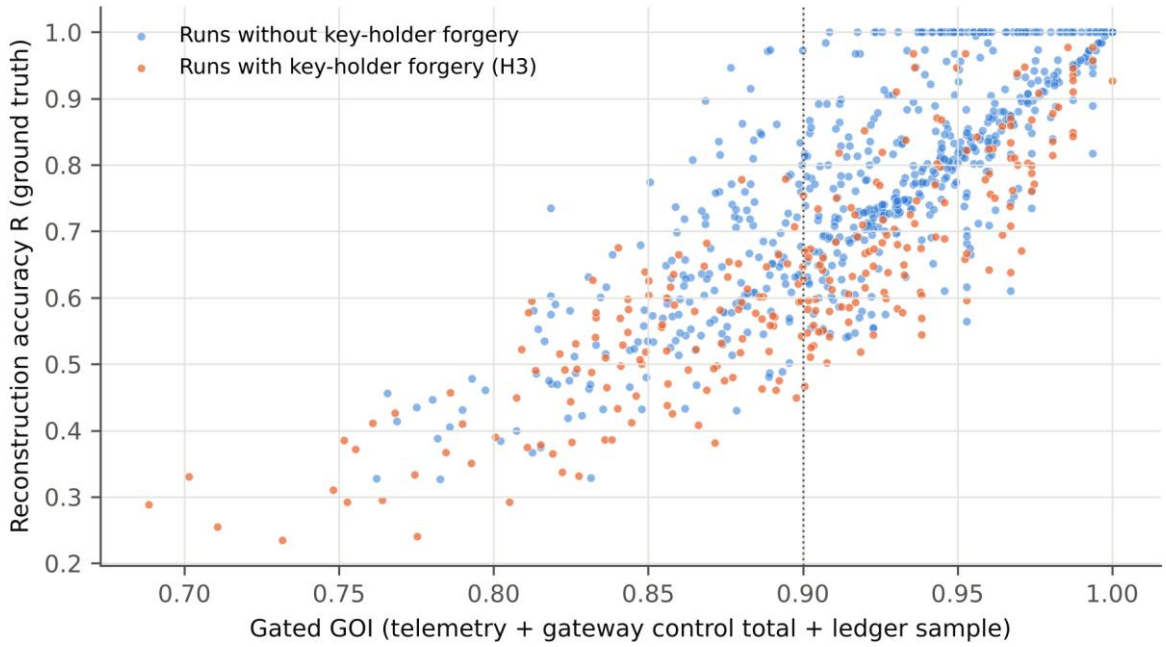


Figure 4. Gated GOI, computed from telemetry together with the gateway control total and the ledger sample, against ground-truth reconstruction accuracy for the 1,000 mixed runs. Orange points contain key-holder forgery.

Figure 4 shows the spread behind the correlation. Of the 703 runs scoring GOI ≥ 0.90, 63.3% were degraded, with reconstruction accuracy as low as 0.467 and silent misstatement as high as 0.364; 60.2% missed at least one episode in time.

5.7 Alarm thresholds and ablations

Used as an alarm, the index needs a threshold much closer to 1 than intuition suggests (Table 5). An alarm at GOI < 0.90 caught only 39.6% of degraded runs. The Youden-optimal threshold in these data was 0.9675, equivalent under compression to a 12% loss in a single dimension. This value is exploratory and depends on the mixture distribution and degraded label used here; it is not a deployment threshold, which these data cannot establish. A rule that alarms when any single dimension falls below 0.90 reached the same sensitivity (0.932) at slightly lower specificity, and it names the failing dimension.

Table 5. Alarm rules on E2 (positive = degraded run). Thresholds are exploratory and specific to this testbed.

Rule	Sensitivity	Specificity	Share of runs flagged
GOI < 0.90	0.396	0.981	0.297
GOI < 0.95	0.810	0.840	0.639
GOI < 0.9675 (Youden optimum, exploratory)	0.932	0.741	0.755
GOI < 0.98	0.972	0.627	0.814
Any dimension < 0.90	0.932	0.703	0.765
Any dimension < 0.95	0.976	0.574	0.831

The ablations answer two different questions, reported separately. Table 6a holds the criterion fixed at the full verifier's outcomes and asks how well each ablated score ranks runs against it. Table 6b recomputes R, U and the degraded label from the records each ablated verifier itself accepts, and asks what that verifier's reconstruction is worth. Removing the control total or the ledger sample changes the score but not which records are accepted, so their Table 6b outcomes equal the full verifier's. Removing run binding and the hash chain changes acceptance itself.

Table 6a. Ablations on E2 against a fixed criterion (the full verifier's R, U, timely recall and degraded label). AUC refers to the degraded label only.

Estimator variant	ρ with R	ρ with 1 − U	ρ with timely recall	AUC
Full	0.849	0.454	0.618	0.925
Without gateway control total	0.797	0.499	0.636	0.892
Without ledger sample ($\hat{F} = 1$)	0.558	0.001	0.454	0.822
Without run binding and hash chain	0.847	0.453	0.611	0.926

Table 6b. Reconstruction produced by each verifier, with R, U and the degraded label recomputed from its own accepted records; E2, n = 1,000.

Verifier	Mean R	Mean U	Mean U on replay runs	Degraded	ρ(score, own R)	AUC (own label)
Full	0.758	0.073	0.087	737	0.849	0.925
Without gateway control total	0.758	0.073	0.087	737	0.797	0.892
Without ledger sample	0.758	0.073	0.087	737	0.558	0.822
Without run binding and hash chain	0.722	0.105	0.212	795	0.889	0.962

Against the fixed criterion (Table 6a), the ledger sample is the most consequential input. Without it, correlation with R fell from 0.849 to 0.558 and correlation with 1 − U fell to about zero. Removing the control total lowered AUC from 0.925 to 0.892. Removing binding barely changed ranking against the fixed criterion.

Table 6b shows why ranking performance and evidence quality must be reported separately. Without binding and chain checks, replayed records were accepted, mean silent misstatement rose from 0.073 to 0.105 (from 0.087 to 0.212 on runs containing replay), and the number of degraded reconstructions rose from 737 to 795. Yet the same variant's score discriminated better against its own degraded label (ρ = 0.889; AUC = 0.962), even though the reconstructions it was judged on were worse in absolute terms. A score can rank outcomes well while the process that produces them accepts more false evidence, so a higher AUC for an altered verifier does not mean it yields safer evidence.

The record counts show the mechanism. On the 250 mixed runs containing replay, removing binding raised the mean number of accepted records from 684.2 to 800.4. But correct reconstructed transfers fell from 144.6 to 128.9, false ones rose from 13.8 to 34.7, and true transfers missing from the reconstruction rose from 40.8 to 56.5. Correct reconstructions fell, rather than staying level, because of how the auditor reconstructs: it keeps one outcome record per action identifier, and a replayed record from another run carries identifiers that also occur in the current run. Once accepted, a replayed record can displace the genuine record for its action, so a false claim replaces a true one. Accepting more records therefore produced a worse reconstruction, not a fuller one.

A further sensitivity analysis tests the fidelity definition. If sampled actions without a valid record are counted as mismatches (an unconditional fidelity over all 40), the index's correlation with R falls from 0.849 to 0.823 and its AUC rises from 0.925 to 0.948. The AUC gain comes from counting missing evidence twice: the unconditional estimate correlates with $\hat{C}$ at $\rho = 0.667$, against -0.025 for the conditional estimate, and its correlation with $1 - U$ falls from 0.971 to 0.486. The conditional definition is retained because it keeps coverage and fidelity separable and diagnostic—the same objection Section 3.2 raises against integrity entering the index twice—subject to the qualification in Section 5.8.

5.8 Forgery and deletion together

Because key-holder forgery and deletion both prioritise structuring transfers, and forgery is applied first, deletion in E2 preferentially removed forged records. In the 66 mixtures containing both faults, 791 of 1,915 forged committed transfers (41.3%) were deleted, against 1,270 of 10,358 unforged committed transfers (12.3%). Records were therefore not missing at random with respect to falsity in these runs. E2b repeats the same mixtures with deletion chosen independently of the forgery targets (Table 7).

Table 7. The 66 mixed runs containing both key-holder forgery and deletion, under targeted (E2, as reported) and independent (E2b) deletion. Falsification rates include emitter-bug records where F2 is also present.

Quantity	Targeted deletion (E2)	Independent deletion (E2b)
Forged transfers deleted	791 / 1,915 (41.3%)	331 / 1,915 (17.3%)
Unforged transfers deleted	1,270 / 10,358 (12.3%)	1,677 / 10,358 (16.2%)
True falsification rate, all executed transfers	0.195	0.195
True falsification rate, transfers with an available record	0.153	0.197
Estimated mismatch rate, $1 - \hat{F}$	0.161	0.199
Mean $\hat{C}$	0.776	0.775
Mean U	0.153	0.197
Mean R	0.590	0.546
Mean GOI	0.866	0.857

Quantity	Targeted deletion (E2)	Independent deletion (E2b)
Timely monitor recall	0.000	0.025

The conditional fidelity estimator tracked the falsification rate among available records in both cases. Under targeted deletion, that rate (0.153) understated the falsification rate across all executed transfers (0.195): deletion converted part of the fidelity defect into a coverage defect. Coverage barely changed, because the deletion volume was the same; what changed was which records disappeared. Silent misstatement was lower under targeted deletion only because more of the false records had been removed rather than accepted, and timely detection was lost in both conditions. The joint runs were degraded under either mechanism: 64 of 66 under targeted deletion and all 66 under independent deletion. Across all 1,000 mixtures, the change made no material difference to the headline results (E2b: ρ with R = 0.851, AUC = 0.923, 738 degraded runs, against 0.849, 0.925 and 737 in E2). Resampling the 1,000 seed pairs with the pairing preserved (2,000 replicates), the differences (E2b − E2) were 0.002 in ρ with R (95% bootstrap CI −0.001 to 0.005) and −0.002 in AUC (−0.006 to 0.001). Both intervals include zero. The paired comparisons therefore give no evidence of a material change in the headline discrimination measures under independent deletion. This is consistent with the principal results being robust to the deletion mechanism, although it does not establish formal statistical equivalence. Mean R and mean U did differ, by −0.002 (−0.003 to −0.001) and 0.003 (0.002 to 0.004): the deletion mechanism changed the quality and composition of the remaining evidence even where the headline discrimination measures did not move. The conditional fidelity estimator should therefore be read as fidelity of the evidence that remains. Extending it to all executed actions requires an assumption about which records disappear, and a key holder who can delete selectively can violate that assumption.

6. Discussion

6.1 What the index measures

Under these simulated conditions, this implementation of GOI is a reasonable screening measure of eventual reconstructability. It correlates strongly with ground-truth reconstruction accuracy, tolerates benign infrastructure behaviour that defeats a strict conjunction of checks, and detects structure-preserving attacks that the legacy conjunction misses—provided it is fed by evidence channels independent of the telemetry stream. That proviso carries most of the weight. Telemetry cannot certify its own completeness or its own truth. The two most consequential estimator inputs in this study, the gateway control total and the ledger sample, are exactly the two that do not come from the stream. This matches the audit principle that external evidence outranks internally

generated evidence (IAASB, 2009a). A deployment that computes GOI from telemetry alone will measure the self-consistency of its evidence, not its reliability.

One might object that an operator able to reconcile against the ledger has no need of telemetry. The ledger, however, records effects only. The governance-relevant state also includes the intent, the authority invoked and the policy decision, which exist only in telemetry; sampling the ledger is a cheap way to test whether that richer record can be believed.

6.2 The gate and the failure that deceives

The rationale for the integrity gate is that untrustworthy evidence must not be masked by completeness. In this testbed the implemented gate did little of that work. It was inactive across almost the entire fault range, because it waits for half the records to fail verification. And where verification works, integrity failures are safe failures: a record that fails its MAC is rejected and becomes a gap, which lowers reconstruction but does not deceive it. The deceptive failure—a false record accepted as genuine—came from evidence that passed every integrity check, because an emitter recorded it wrongly or a key holder re-signed it.

Within the checks implemented in this study, ledger reconciliation provides the direct test of whether the recorded effect matches the executed transfer. It is not the only possible signal. Where a falsification touches some spans of an action but not others, the spans disagree, and a cross-span consistency check can reveal the inconsistency without the ledger, though not which value is true. Applied to the authenticated records, such a check flagged every run containing the emitter bug (F2, which alters the tool and outcome spans only) at both $p = 0.05$ and $p = 0.30$, and flagged no clean run. It flagged none of the single-family forgery runs in v1.1, where the forger rewrites all four spans consistently. The v1.0.1 forger left the policy-decision span unchanged; against it, the same check flags all 180 single-family forgery runs at $p > 0$. Consistency checks are therefore cheap and worth adding, but they catch careless falsification rather than a competent key holder.

The design implication is that if any dimension deserves a non-compensatory gate, it is fidelity, and the gate should act on a confidence bound rather than a point estimate. Consider a run in which all 40 sampled transfers are present and match. Under a binomial model, which treats sampling as with replacement from a large population, the one-sided 95% upper bound on the mismatch rate is 7.2%. Under the hypergeometric model that matches this testbed—40 drawn without replacement from 177–190 committed transfers—it is about 6%. In either case the bound is not zero, and it widens as n_F falls. Integrity should remain a precondition for admitting records to the reconstruction, which the estimator already enforces record by record; it adds little as a multiplier on the score. These recommendations come from one testbed with one threat mix and should be tested before adoption, not treated as calibrated.

6.3 Compression and thresholds

The fourth root crowds the useful range into [0.9, 1.0], and in these data a threshold of 0.90 missed most degraded runs. Two remedies are available without changing the construct. One is to report the four-dimension vector with per-dimension floors as the decision rule, using the composite for trend and ranking. The other is to rescale the composite (for example, the product $C \cdot F \cdot I \cdot T$ without the root) so that equal losses produce visibly equal drops. The reasoning that rejects a single weighted score as the gate for an assurance decision applies equally to the GOI itself. Whichever form is reported, it should carry an uncertainty statement. The four estimates have different error sources—$\hat{C}$ depends on the accuracy of the control total, $\hat{F}$ on a sample of up to 40 with an interval several points wide, and $\hat{T}$ on clock synchronisation—so a point score alone overstates what is known.

6.4 Not all records are equal

Reconstruction accuracy weights every committed transfer equally. Governance does not. Targeted deletion of 5% of actions—the structuring transfers first—erased timely detection, and targeted forgery of 5% nearly did so (timely recall 0.022), while the index barely moved, because coverage and fidelity count records without regard to their governance weight. A risk-weighted coverage measure, counting records from high-consequence actions or open monitor windows more heavily, would close part of this gap. The broader point is familiar from the measurement literature: once an index gates decisions, a motivated actor will degrade whatever the index weights least (Campbell, 1979). Section 5.8 shows the same logic across dimensions: a key holder able to delete as well as forge can move a defect from the dimension that would expose it to one that only registers a loss.

6.5 Relation to assurance practice

A governance observability score is evidence inside an assurance argument, not a substitute for one. Safety-case practice has a documented failure in which the argument became a paper exercise detached from the hazards it claimed to address (Haddon-Cave, 2009). The equivalent risk here is an index reported upward as a single reassuring number. The results argue for reporting GOI together with its evidence sources (control totals, sample sizes and confidence bounds), the threat classes it can and cannot see, and the distinction between eventual reconstruction and timely warning, so that a reader knows what a score of 0.99 does not rule out.

7. Threats to validity

Construct validity. The criteria—reconstruction accuracy, silent misstatement and timely recall of one structuring pattern—are the author's operationalisation of reconstructability. Other reasonable

criteria, such as explanation of intent or provenance of retrieved context, are not tested. The original proposal names four evidence classes—decision traces, tool-call records, context provenance and intervention logs. Decision traces and context provenance are absent here; the testbed scores transactional records only. Two of the reported associations are partly definitional ($\hat{C}$ with R under random loss; $\hat{F}$ with $1 - U$), as Section 5.6 explains.

Internal validity. The same author built the estimator, the faults and the criteria. The out-of-specification families were not blind, and testbed defects were corrected after smoke tests (Section 4.5). The v1.1 change to the forgery fault was made after peer review identified that the earlier version left one span unaltered. The released code allows each of these choices to be varied; independent fault generation by a third party is the next step.

External validity. The agent is a scripted generator, not a language model. The evidence-quality questions studied here are not specific to agents and also arise in other automated systems that emit signed telemetry. The study therefore examines the evidence-quality layer that agent governance depends on. It does not test what makes agents harder to observe: goals that must be inferred, an open action space, and delegation across trust boundaries. The domain is a single payments workflow; the fault distribution is assumed, not measured; and severities are uniform within families. Real telemetry has correlated failures, partial instrumentation and ambiguous authority, none of which is modelled. The control total is assumed error-free. Results such as the AUC and the exploratory threshold depend on the mixture distribution and should not be quoted as operational performance. The key-holder adversary is a specific one: it can forge and delete telemetry but cannot alter the gateway control total or the ledger. An insider who controlled either independent channel as well could defeat the checks that caught forgery and deletion here. The sequence-level criterion is a single structuring rule; other violation patterns may depend on different records.

Statistical conclusion validity. All intervals are bootstrap intervals over simulated runs. They describe uncertainty about the simulation's own behaviour, not about any deployment. Per-run fidelity estimates rest on between 9 and 40 compared records (574 of the 3,520 E1 and E2 runs had fewer than 30), so their precision varies with coverage.

Missingness. Conditional fidelity describes the evidence that remains; it equals fidelity across all executed actions only if records go missing independently of whether they are false. In the single-family experiments that holds, because each run has one fault. In the mixed experiment it does not hold for runs containing both key-holder forgery and deletion, where deletion preferentially removed forged records (Section 5.8). In this testbed the control total registers the resulting loss as

reduced coverage. A deployment without an independent count would not see it, and a deployment whose control total the key holder can also alter would not either.

8. Conclusion

The earlier experiments showed only that a binary verifier implemented its own checks. This study asked whether one implementation of the continuous Governance Observability Index tracks how well an agent's governance state can be reconstructed. In a scripted simulation with the complete ledger and fault labels withheld from the estimator, it largely does for eventual reconstruction (ρ = 0.85; AUC 0.93), less well as a warning of missed time-sensitive violations (AUC 0.87), and it tolerates benign infrastructure behaviour that defeats a strict check-list verifier. The qualifications matter as much as the headline. The index's validity here comes mostly from evidence channels outside the telemetry stream. Its fourth-root scale makes 0.90 a lenient threshold. In this testbed, the implemented integrity gate guarded a failure that verification already made safe. The failure that deceived—validly signed false evidence—passed through fidelity, and a key holder able to delete as well as forge could shift part of that failure into coverage.

Within the limits of a scripted simulation, these results support using GOI as a screening and trend measure, reported with its dimension vector and evidence sources, not as a certificate that an agent's record can be trusted. The next steps are a fidelity gate based on a confidence bound, risk-weighted coverage, cross-span consistency checks, independently generated faults, and a live-model sandbox in which the four dimensions are measured from real agent traces.

Declarations

Funding. The author received no support or funding from any organisation for the submitted work.

Competing interests. The author proposed the Governance Observability Index evaluated in this study. The author reports no commercial engagement with any provider of AI-governance products or services and no other competing interests.

Ethics approval and consent to participate. Not applicable. The study used simulated data only; no human participants, personal data or real financial transactions were involved.

Author contributions. Ravi Shankar NRK conceived the study and its research questions, directed the design of the testbed and analyses, reviewed the released code, its outputs and the record of its independent re-execution, checked the reported results against those outputs, prepared and revised the manuscript, and takes responsibility for the work.

Use of generative AI. Author has used Claude (Anthropic) for writing and execution of simulation and analysis code (releases v1.0 to v1.1) under his direction, and drafted substantial parts of the

manuscript text and revisions. ChatGPT (OpenAI) and Gemini (Google) were used for adversarial critique, language editing and checking; ChatGPT was also used to re-execute the released code in a separate computing environment and compare its outputs with the archived results. The author checked sources against primary records, reviewed the code and outputs, revised the manuscript and accepts full responsibility for the final work.

Acknowledgements. None.

Data availability

The materials supporting this article are release v1.1 of the simulation package, published on Zenodo (NRK, 2026; https://doi.org/10.5281/zenodo.24305314) and supplied with this article as Online Resource 1. The release comprises the simulation and analysis scripts, the run-level dataset (4,520 rows: E1, E2 and E2b), the run manifest, summary outputs, Tables 6a, 6b and 7 as CSV files, figures, pinned analysis dependencies, a record of an independent re-execution, and SHA256SUMS.txt with a SHA-256 checksum for every file. All columns of release v1.0.1 are retained, and the per-ablation outcome columns used in Table 6b are included. The materials reviewed with the first submission (release v1.0.1) remain unchanged in the earlier Zenodo deposit (https://doi.org/10.5281/zenodo.23936145); that archive's README stated that no DOI had been assigned, which was incorrect.

To reproduce the results without overwriting the archived outputs, run `python3 goi_testbed.py repro` followed by `python3 analyse.py repro` from the package directory, and compare the new folder with the archived results folder. The outputs reported here were produced with Python 3.13.18, NumPy 2.5.3, pandas 3.0.5, SciPy 1.18.1 and Matplotlib 3.11.2, under which the rerun reproduces the dataset and tables byte for byte; the testbed itself uses only the Python standard library. The master seed is 20261029, and the signing keys are test-only literals. Online Resource 2 tabulates the definition of the index as originally proposed and as implemented in this study.

References

Campbell, D. T. (1979). Assessing the impact of planned social change. Evaluation and Program Planning, 2(1), 67–90.

Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., Bucknall, B., et al. (2024). Black-box access is insufficient for rigorous AI audits. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 2254–2272). ACM.

Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., et al. (2023). Harms from increasingly agentic algorithmic systems. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 651–666). ACM.

Crosby, S. A., & Wallach, D. S. (2009). Efficient data structures for tamper-evident logging. In Proceedings of the 18th USENIX Security Symposium (pp. 317–334). USENIX Association.

Dong, L., Lu, Q., & Zhu, L. (2024). AgentOps: Enabling observability of LLM agents. arXiv preprint arXiv:2411.05285.

Haddon-Cave, C. (2009). The Nimrod Review: An independent review into the broader issues surrounding the loss of the RAF Nimrod MR2 aircraft XV230 in Afghanistan in 2006 (HC 1025). The Stationery Office.

International Auditing and Assurance Standards Board. (2009a). International Standard on Auditing 500: Audit evidence. International Federation of Accountants.

International Auditing and Assurance Standards Board. (2009b). International Standard on Auditing 530: Audit sampling. International Federation of Accountants.

Jia, Y., & Harman, M. (2011). An analysis and survey of the development of mutation testing. IEEE Transactions on Software Engineering, 37(5), 649–678.

Kalman, R. E. (1960). On the general theory of control systems. In Proceedings of the First International Congress of Automatic Control (pp. 481–492). IFAC.

Leucker, M., & Schallhart, C. (2009). A brief account of runtime verification. Journal of Logic and Algebraic Programming, 78(5), 293–303.

Messick, S. (1995). Validity of psychological assessment. American Psychologist, 50(9), 741–749.

Moreau, L., & Missier, P. (Eds.). (2013). PROV-DM: The PROV data model. W3C Recommendation.

Munda, G., & Nardo, M. (2009). Noncompensatory/nonlinear composite indicators for ranking countries: A defensible setting. Applied Economics, 41(12), 1513–1523.

NRK, R. S. (2026). Governance Observability Index v 1.1 [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.24305314

OECD & European Commission Joint Research Centre. (2008). Handbook on constructing composite indicators: Methodology and user guide. OECD Publishing.

Reuel, A., Bucknall, B., Casper, S., Fist, T., Soder, L., Aarne, O., et al. (2025). Open problems in technical AI governance. Transactions on Machine Learning Research.

Schneier, B., & Kelsey, J. (1999). Secure audit logs to support computer forensics. ACM Transactions on Information and System Security, 2(2), 159–176.

Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O'Keefe, C., Campbell, R., et al. (2023). Practices for governing agentic AI systems. OpenAI.

Sigelman, B. H., Barroso, L. A., Burrows, M., Stephenson, P., Plakal, M., Beaver, D., Jaspan, S., & Shanbhag, C. (2010). Dapper, a large-scale distributed systems tracing infrastructure (Technical Report dapper-2010-1). Google.

Sridharan, C. (2018). Distributed systems observability: A guide to building robust systems. O'Reilly Media.

United Nations Development Programme. (2010). Human development report 2010—The real wealth of nations: Pathways to human development. Palgrave Macmillan.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems, 12(4), 5–33.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- OnlineResource1GOIvalidityv1.1.zip
- OnlineResource2GOIDefinition.docx