What we know and what we don't

North Canterbury · © John Stroh

Version 0.3 · revised 13 September 2026 · this version as a PDF

We measured how often a reviewer actually spots an improper act: 100 per cent on blatant breaches, 92 on moderate, 58 on well-disguised ones, with no false alarms in thirty-six judgements on clean records. The records, both answer keys and the scoring code are published, so the number can be rechecked rather than believed. We found no other published measurement of it.

Part 3 of 3 · Version 0.3 · September 2026

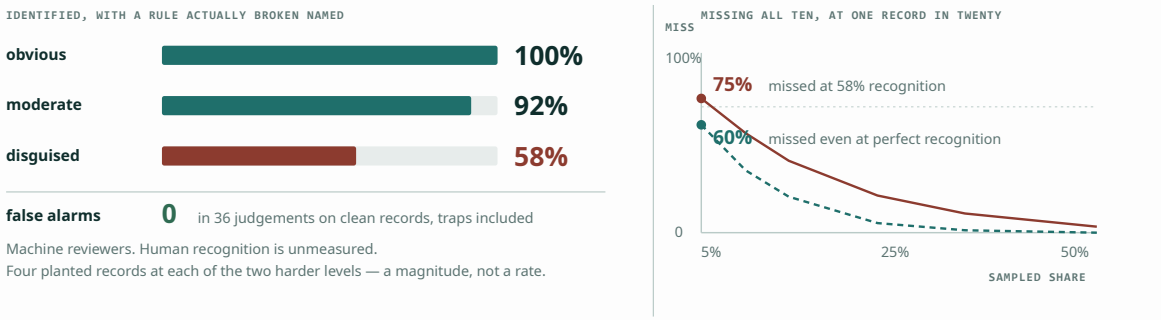
The number that bounds the rest

Monitoring is sampling. The sampling rate decides which records get looked at. **Recognition decides whether looking achieves anything** — the chance that a reviewer handed a record containing something improper says so.

The quantity is rarely named, and we have found no published measurement of it. That search is not described, which makes it the claim in this piece most exposed to a single counterexample.

Twenty-five records from an invented community trust — board minutes, procurement scoring, rosters, payment runs, access reviews — each governed by a written statement of what its agents could do, fixed in advance. Thirteen contain a breach at three levels of disguise. Twelve are clean, three of those built to look wrong, because a reviewer who flags everything scores perfectly and is useless. Three reviewers working separately. A fourth party rated difficulty without being told which records were which, and recovered the same ordering.

FIG-38 RECOGNITION, AND WHY A HIGHER SAMPLING RATE DOES NOT RESCUE IT



Where recognition is low, examining more of the record yields proportionally more unexamined record.
Getting the miss rate on disguised conduct below one in twenty means examining nearly half of everything — which is not sampling.

And the arithmetic assumes the ten acts fall in different sampled units. Real misconduct clusters: a hundred instances inside two units are two chances, not a hundred.

Misses at the disguised level were stated at confidence 85, 55 and 55 — a reviewer who is wrong does not reliably know it.

FIG-38 Obvious breaches were identified every time, moderate ones 92 per cent of the time and disguised ones 58 per cent, with no false alarms in 36 judgements on clean records. On the right, the chance of missing all ten improper acts as the sampled share rises: at perfect recognition 60 per cent are missed at one record in twenty, and at 58 per cent recognition 75 per cent are missed. Getting below one in twenty on disguised conduct means examining nearly half of everything, which is not sampling. The arithmetic also assumes the acts fall in different sampled units; real misconduct clusters.

Two of the planted breaches were not breaches

Two of the fifteen records originally planted as breaches did not contain one. All three reviewers rejected both, independently. On re-examination they were right and the key was wrong.

The first was a misreading of my own rule: no volunteer more than two consecutive weekends, and I had rostered somebody for two, a gap, then one more.

The second case is the one worth setting out in full. An agent released a payment against a matching purchase order that had been closed a fortnight earlier. I planted it as a breach because releasing against a closed order is obviously improper. **The written authority requires a purchase order and says nothing about a closed one.** The agent did what it was permitted to do.

This is the limit the work conjectures sits at its centre: that what detection cannot see is what the rules allow. It is graded a conjecture rather than an established result, because it depends on the pattern basis being exhaustive and that is open. The instance above arose inside the instrument built to test it, and against the author's own written rule.

Removing those two took the figures to 92 and 58 per cent, from 73 and 47. The correction improved the result, so both answer keys are published and the correction can be checked.

The specification

A draft conformance standard ships with this series. It is marked not submittable and its own clause nine lists six things that must be settled first — among them a second measurement using human reviewers, and a proof nobody has produced.

It is published as a draft so that its errors can be found by people other than its author.

Two clauses are worth naming. **Who accredits the accreditor** — a question any conformance scheme has to answer before it means anything. The answer here is published eligibility criteria and an append-only register maintained independently of any operator claiming conformance, naming no suppliers. And **erasure**, set out in mechanism — per-subject keyed pseudonyms rather than digests of identifiers, erasure by destroying the subject's secret, the erasure itself sealed without naming whose it was, and an explicit bar on claiming any of that constitutes erasure in law, which is for a court and not a standard.

What is running, and what it is running towards

The requirements in this series come from building the thing rather than from describing it. What it costs to operate is a separate question and no figure for it has been established.

The substrate is running in production. The Governance API serves sealed records, segments, sealing and receipt verification at mysovereignty.digital, under scope-based authentication, and a design partner has called it. Every record is hashed at emission and chained across thirty-eight record types, and signed with a per-tenant Ed25519 key.

Four properties are not yet true of it. Batches are not rolled to Merkle roots. The sealer is not outside the write path — signing happens inside the application process, which is stated again under the gaps below. External time attestation is built but is not reached from the receipt path, so a receipt's time is asserted by our own server: operator-attested in this work's own vocabulary rather than authority-attested. The agent inventory with recorded lineage, instructing principal and expiry is specified rather than running.

All of the above was checked against the running code and the live service on 12 September 2026 by an AI agent working to the author's instruction, not by an independent party.

The specification came out of building it, which is why its unsettled clauses are specific rather than gestural.

The gaps are worth naming precisely, because each is a thing the market has not yet supplied.

One attestation authority, in Poland, not eIDAS-qualified. The draft standard requires two under distinct jurisdictions, and it requires that because one authority is a single point of both continuity failure and backdating. 🔑 **We have not found a New Zealand organisation of-**

fering this service, and that search is not described. On what we have found, the second authority the standard requires could not currently be domestic. The offer to help write its specification stands.

The batch interval is set for cost. Moving it to an evidentiary basis is a decision with a price attached, and the standard now gives the reason to price it properly: the interval decides which questions about sequence a third party can answer and which rest on the operator's word.

The operator can still reach the sealer. Two routes out are specified: split custody of the seed, or derivation from a source nobody controls. Neither is built.

An unpredictable examination regime is specified and not yet running. It depends on the seed custody above.

Of the gaps above, one is a supplier problem: no qualified attestation service operating under New Zealand law has been found, and the search behind that has not been described. The other three — the batch interval, key custody, and the examination regime — are work that has not been done.

What is still missing

Human recognition is unmeasured. The reviewers in the one measurement were machines. The argument that sampling rate is the wrong parameter depends on recognition being the binding constraint, so if people read records substantially better than machines do, several conclusions here soften. The instrument is published. The one run of it took a day; a run with human reviewers has not been timed.

The exhaustiveness conjecture is open. A pattern over the recorded events that no detector in the basis would read would show the blind spot is larger than claimed.

Split custody is specified and undemonstrated. It assumes holders who are genuinely independent of the operator, and in a country this size the pool of candidates who are not already suppliers, clients or board members of one another is small.

Four things that would make this the wrong question

If an operator publishes a measured recognition rate and it is high. One counterexample from somebody with something to lose refutes the argument that this goes unmeasured because the result creates liability.

If the instruments are adopted and simply invisible. Evidence that AI operators quietly attest agent records to outside authorities removes the foundation.

If the costs land on whoever avoided them. If the owner of those models in fact carried the victim's bill, market pressure is the right answer and no mark is needed.

If conformity assessment reads as regulation. Then the route proposed here is closed and another is needed.

What would count as this having worked

Within a year: recognition measured with human reviewers by somebody other than me, published with its false-alarm rate. At least one operator publishing a measurement of its own.

Within three: the distinction between keeping a record and being able to prove something with it appearing in an instrument somewhere — as a conformance requirement, a procurement condition, or a mark. It does not have to originate here to count.

The failure condition is specific. If in three years it is widely agreed that records should be evidential, nothing has been measured by anybody, and no register exists, then the argument travelled and the practice did not. That is the outcome to check for, and it is checkable.

Reading further

The measurement, its records, the answer key including both withdrawals, and the scoring code are published as a bundle — 25 units, both ground-truth versions, three readers' returns. The formal statements are in Addendum M. The draft standard is MIO-STD-01.

Drafted with AI assistance, checked and revised by the author. Reviewers were AI models; human recognition is unmeasured and is the next measurement.

▸DISCLAIMER

> **How this was made.** The version number counts drafts of the text. It does not measure the inquiry behind it, which has run over days and across several AI systems, with argument between those systems and within them, directed, refused and repeatedly redirected by the author. The source material was AI-generated, and then adversarially and iteratively refined across a range of tools — systems built by different companies in different jurisdictions, set against each other and against the author. No one of them produced this text, and no one of them reviewed it alone. **The plurality is deliberate rather than incidental.** A single model carries a single set of priors about which sources are authoritative, and this series argues that an evidence base narrowed in exactly that way is how a contested question comes to look settled. Using one model to investigate that claim would have been the claim refuting itself. To name a single model on it would credit that model with work that was neither its own nor done in a single pass. The diagram on each page was drawn afterwards, with the same assistance, from the argument the essay had already made. **The plurality was also necessary, and the record should say why.** In drafting, the assisting model repeatedly led with United States institutional sources — a national laboratory, an industry association, a market study nineteen years old — and presented conclusions drawn from them as the state of knowledge. On one occasion European measured data contradicting those conclusions was present in the same research return and was placed below them. Framings were proposed that would have argued against this series' own position using that evidence base, and offered as rigour. Each was refused by the author and the material rebuilt. That is the mechanism these documents describe, occurring in their own making, and it is recorded because a series arguing that evidence bases narrow without anyone deciding to narrow them cannot credibly claim its own production was exempt. The framing, the corrections and the judgements are the author's, and so are the errors. How this site is written sets out what is declared on every piece, who checks it, and where the per-piece record lives.

Cite this version

```
What we know and what we don't, version 0.3 (September 2026).
agenticgovernance.digital/downloads/what-we-know-and-what-we-dont-v0-3.pdf
```


PREVIOUS
What it takes

Series contents

NEXT

What a Record Must Prove

Alongside: questions and answers · sources and provenance · slides