What we claim, and how sure we are

Aotearoa New Zealand · © John Stroh

Version 0.4 · revised 13 September 2026 · this version as a PDF

Everything this work asserts, with a grade on each: what follows from standard results, what is stated precisely enough to be tested but is not proved here, and what is an open question we have not answered. Every claim says what would show it to be wrong.

Claims register · Version 0.4 · September 2026

The claim

If the only copy of what your AI systems did is the one you hold, nobody outside your company can check it.

An organisation running AI agents can meet every record-keeping duty in force and still be unable to show anybody else what its agents did. The instruments that change that have been standard for twenty-five years, and we have not found them in use for this purpose. This work sets out what using them takes: built, measured once, and written up as a draft standard with its unsettled questions listed.

How to read the grades

GRADE	MEANS
Established	Follows from standard results. No novelty is claimed; the contribution is the composition.
Conjectured	Stated precisely enough to be tested. Not proved here.
Open	A question this work poses and does not answer.

A claim that depends on a conjecture cannot itself be established. E5 is absent from the established list below for that reason: it was reclassified as C4 on 12 September 2026.

What is established

	CLAIM	WHAT WOULD SHOW IT WRONG
E1	The substrate is a filtration: the past is annotatable, never editable, and a detector at any moment sees everything sealed before it.	A construction that alters a sealed record without breaking the chain.
E2	Commit-and-reveal inherits its unpredictability from the external clock alone. Without one, the committing party controls the order and the proof collapses.	A commitment scheme that is unpredictable to its own author without an outside time source.
E3	The batch interval Δ is a parameter of the evidence, not of cost. Every statistic carries an attestation grade.	Showing that events inside one batch can be independently ordered.
E4	The null-source trichotomy — learned, authored, randomised — is exhaustive for detectors whose inputs are the sealed history, the mandate and the seed.	A fourth kind of null model built from those inputs.
E6	A single fixed examination configuration can be driven towards zero detection by an adversary who can identify it. A mixture over a closed set has a floor.	A fixed configuration with a floor, or a mixture without one.
E7	Deriving a separate stream per component gives independence to an observer and joint reproducibility to an auditor.	Recovering one component's stream from another's observed behaviour.
E8	Among point processes of a given rate, homogeneous Poisson timing is the maximum-entropy choice: an observer learns least from it.	A timing distribution at the same rate from which an observer can infer more about the active configuration.

What is conjectured

	CLAIM	WHAT WOULD SETTLE IT
C1	A learned null's detection power degrades at least linearly in the fraction of history that is already compromised.	Proof or refutation for a specific estimator.
C2	The eight pattern types span the patterns definable over the substrate with first-order statistics on labels.	A pattern outside the basis. This is the load-bearing conjecture; C4 rests on it.
C3	At governance scale a near-uniform mixture is within a small constant of the optimal one.	Computing the optimum for a real rule set.
C4	The residual — what no detector can see — is exactly the set of acts the written authority permits, which is to say mandate error.	**Downgraded from established, 12 September 2026.** The identity holds only if C2 does. What is not in question is the weaker half — no detector flags an act the written authority permits, because there is nothing to flag against. Whether that set is *exactly* the residual depends on C2.

What is open

	QUESTION
O1	Compute the detection floor for a concrete rule set and adversary class. It is a number rather than a principle, and we have not produced one.
O2	Quantify what the content of released findings leaks about the active configuration.
O3	The practical cost of spreading a small examination budget across detectors.
O4	Whether C2 holds. If it does not, the basis needs extending and C4 goes with it.
O5	The recognition rate enters the detection probability multiplicatively and is unmodelled in the formal work.
OQ-15	**What does any of this detect when an agent has no owner?** Recorded 12 September 2026. The external clock and the unpredictable seed survive an ownerless agent. Sealer isolation survives only where there is an estate to run one. The fourth — a written authority somebody answers for — does not exist, and with no mandate the residual is everything.

The one measurement, what it shows and what it does not establish

Recognition rate *q*: the probability that a reviewer handed a record containing an improper act says so. It matters because sampling and recognition **multiply** — the chance of missing all *n* improper acts is (1 − s·q)^n for sampled share *s*, so a perfect sampling regime staffed by reviewers who do not recognise what they see catches almost nothing.

Measured once: **100, 92 and 58 per cent** across three levels of disguise, with no false alarms in thirty-six judgements on clean records. The records, both answer keys and the scoring code are published as a bundle.

It does not establish a threshold. The reviewers were machine reviewers; human recognition is unmeasured, and that is the largest open item in this work. Twenty-five units, with five planted breaches at the blatant level and four at each of the others after withdrawal, gives a magnitude — that recognition is neither near one nor near zero — and nothing more. The units are synthetic.

Two units planted as breaches were **withdrawn during scoring**. All three reviewers rejected them independently; on examination the units did not contain what the key said they contained. That moved two figures up, from 73 and 47 per cent. Both keys are in the bundle, so the correction can be checked.

The standard is a draft and says so

MIO-STD-01 is published at working-draft status. Its clause 9 lists what must be settled before it could be submitted anywhere:

1. *q* has been measured once, for model readers only — not enough to set a threshold.
2. The exhaustiveness conjecture (C2) is open.
3. The delta over PeerReview must be stated precisely.
4. No attestation authority is named, and the standing of a foreign one is unexamined.
5. The intellectual-property declaration differs between standards regimes.
6. Alignment with ISO/IEC 24970 is unexamined.

What is running, and what is not

An implementation of this standard runs in production. Its position against the standard is set out below in both directions. Each row 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.

What is running.

The Governance API is live	Sealed records, segments, sealing and receipt verification, under scope-based authentication, at `mysovereignty.digital`. A design partner has called it.
Every record is hashed at emission and chained	Across thirty-eight record types, with a per-tenant Ed25519 signature over a canonical payload.
Receipts are signed fail-closed	A receipt is never issued unsigned, and never without an explicit, signed label naming which authority reached the verdict. That label is about the decision, not about the time.
Records are anchored to an outside authority, on a daily sweep	The sweep submits record digests to an RFC 3161 authority and stores the returned token whole, with its certificate chain, so a third party verifies it with `openssl ts -verify` and never has to contact us. It runs after the fact rather than at the moment a record is made — see the shortfall on receipt time below.
A partial inventory of credentials is running	Scoped API keys record who minted them, what scopes they carry, when they expire, and when they were revoked or last used; every proof-chain entry records the key, the signer and what evaluated the policy. This establishes which credential acted. It is not the agent inventory the standard requires — see the shortfall below.

What is not. Each is a shortfall against a clause of the standard.

One attestation authority, in Poland, not eIDAS-qualified	The standard requires two under distinct jurisdictions. The authority in use is free and carries no statutory presumption; it is a deliberate placeholder for a commercial authority, and what it supplies today is the evidence rather than the presumption. No qualified service operating under New Zealand law has been found, and that search has not been described.
The batch interval was chosen for token cost	The standard says it is a parameter of the evidence. The implementation does not meet its own standard here, and the interval would have to be reset against the conduct it must attest.
Key custody is undecided	The operator holds the sealer, seed, inventory and observer keys. Two routes out are specified — split custody of the seed, or derivation from a source nobody controls — and neither is built.
The inventory does not record lineage or a sealed mandate	Scoped keys establish which credential acted. They do not establish which agent instantiated which, under whose written authority, or when that authority expires.
A receipt's own time is operator-asserted	The daily sweep anchors records after the fact. The time inside a governance receipt comes from our own server at the moment it is issued: grade O in this work's vocabulary, not grade A. Wiring the attestation into the receipt path is written and not yet deployed.
The unpredictable examination regime is not running	It depends on the seed custody above.

Where to go next

The argument for why this matters: Cheaper not to look. What catches an agent: What it takes. The evidence: What we know and what we don't. How to build it: How to build it. The formal statements: Addendum M. The standard: MIO-STD-01.

Drafted with AI assistance, then checked and revised by the author.

▸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 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 claim, and how sure we are, version 0.4 (September 2026).
agenticgovernance.digital/downloads/what-we-claim-v0-4.pdf
```


PREVIOUS
How to build it

Series contents

NEXT
What was visible at the time