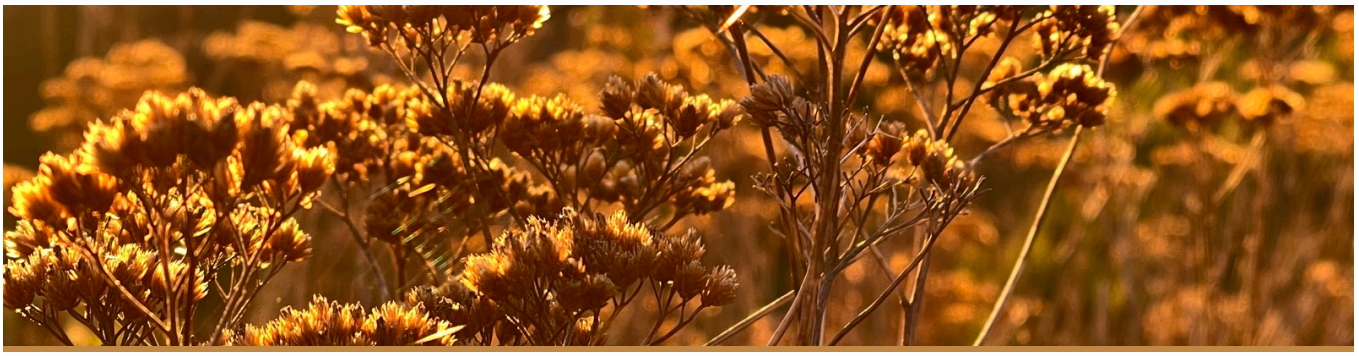

Addendum M — Mathematical Strategies for Timelined Detection

Aotearoa New Zealand · © John Stroh

Version 0.1 · revised 11 September 2026 · this version as a PDF

> Drafted with AI assistance, then checked and revised by the author. The judgements and the errors are the author's.

The formal statements behind the Proof of conduct series. Every claim is graded established, conjectured or open, and the grades are the point: the essays state results in prose, and this is where a reader checks what each one rests on.

Formal addendum · Version 0.1 · September 2026

Addendum to the Sealed-Substrate Audit Construct (v0.2). Drafted 11 September 2026 by Claude Fable 5.1, an AI model, at John Stroh's request. Setting-agnostic: nothing here depends on the Village, on ownership, or on media type.

0. Status of claims

Three grades are used throughout.

- **Established** — follows from standard results (measure-theoretic filtrations, minimax for finite zero-sum games, maximum-entropy point processes, indistinguishability bounds). No novelty claimed; the contribution is composition.
- **Conjectured** — plausible, stated precisely enough to be tested, not proved here.
- **Open** — a question the framework poses but does not answer.

Assumptions the whole framework rests on:

- **A1 External clock.** An authority independent of every actor attests a monotone time T on batches, and cannot be induced to backdate within the batch interval Δ.
- **A2 Unpredictable seed.** A seed σ is fixed before a period and is unpredictable to every actor, including the operator, until reveal.
- **A3 Sealer isolation.** No actor can write to the log or the signing key.
- **A4 Mandate correctness.** The authored null (mandate, inventory, separation-of-duty rules) admits no harmful act. Acts it admits are out of scope for detection by construction (Section 4.4).

A1–A3 are engineering and custody. A4 is governance. The mathematics starts after them.

1. The substrate as an object

A substrate is $S = (E, \prec, \lambda, \rho)$.

- E: a set of events.
- $\lambda: E \to L$, a labelling giving each event an actor a(e), a type τ(e), a target, and a content hash h(e). Content itself is not in S; only h(e) is. This is why the framework is modality-blind.
- $\rho \subseteq E \times E$: reference edges (hash pointers, derivation, delegation, citation).
- $\prec$: a partial order on E, composed of two relations of different provenance:
  - $\prec_A$, authority-attested: $e \prec_A e'$ iff batch(e) precedes batch(e′) under T.
  - $\prec_O$, operator-attested: $e \prec_O e'$ iff e precedes e′ in the same hash chain within one batch.

$\prec$ is the transitive closure of $\prec_A \cup \prec_O$. Every temporal claim made by any detector has an **attestation grade**: A if it needs only $\prec_A$, O if it needs $\prec_O$. Grade-O claims are only as trustworthy as the operator. (Established: this is just bookkeeping, but it is bookkeeping the construct must carry on every statistic.)

2. The role of the timestamp

The timestamp does three distinct jobs. Conflating them is where earlier drafts went soft.

2.1 Direction: the substrate is a filtration

Let F_t be the set of events sealed by authority time t. Because T is monotone (A1) and the log is append-only (A3), $F_s \subseteq F_t$ for $s < t$, and F_t is fixed at t for all later times. This is a filtration in the ordinary sense.

Consequences (established): - Any detector is a process adapted to the filtration: what it can compute at t is a function of F_t and nothing later. - No actor can alter F_t after t. This is the precise content of “the audit is intermittent, the evidence is not.” - Corrections are new events in F_{t′}, t′ > t, referencing the old. The past is annotated, never edited.

2.2 Binding: commitments are meaningful only under an external clock

A commitment c = h(σ) sealed in F_{t0} and a reveal σ sealed in F_{t1}, t0 < t1 under ≺_A, prove that σ was fixed before anything it governs. Without A1 the proof collapses, because the committer controls the order. Every randomised element of the construct — observer schedule, sampling units, batch jitter, release times — inherits its unpredictability guarantee from this single mechanism. (Established.)

2.3 Resolution: Δ is an evidentiary parameter

Two events in the same batch are unordered under ≺_A. So:

- Any statistic requiring order finer than Δ is grade O.
- Density, synchrony and sequence detectors degrade to grade O as their characteristic timescale falls below Δ.
- The batch interval is therefore a parameter of the *evidence*, not of cost, and should be chosen against the timescale of the patterns the deployment must attest independently.

If batch boundaries are predictable, an actor can position acts within a batch to keep its ordering at grade O. Hence batch boundaries are themselves drawn from σ (Section 5.4).

2.4 The timeline’s own signature

The batch sequence B = ((T_1, n_1, g_1), (T_2, n_2, g_2), …), with n_i the batch size and g_i the undated-gap flag, is a time series in its own right, distinct from the events inside batches. Pattern types (Section 3) apply to B: drift in batch size, clustering of gaps, synchrony of gaps with actor activity. This is the “signature of the linear necessity”: a timelined strategy cannot avoid producing a timeline, and that timeline is auditable by the same tools. (Established, once stated.)

3. Detector algebra

3.1 A detector

D = (N, s, θ, U, r, g).

- N: null model — a distribution over benign substrates, or a predicate that benign substrates satisfy.
- s: a statistic on (S restricted to a sampling unit).
- θ: a threshold or decision rule.
- U: the sampling unit (record, window, actor-trajectory, subgraph).
- r: required temporal resolution; determines attestation grade g ∈ {A, O} via r ≷ Δ.

Output: findings, which are events in E with τ = finding, referencing what they judge. Detectors are therefore actors in S. This closes the loop used in Section 6.

3.2 Basis detectors from the taxonomy

The eight pattern types are the basis. Each is a class of D distinguished by the structure s reads:

TYPE	S READS	TYPICAL U	GRADE RISK
Serial	frequency along a trajectory	record	A
Cluster	local density	window	O below Δ
Sequence	ordered typed composition	trajectory	O below Δ
Relational	motifs in ρ	subgraph	A (ρ is hash-bound)
Absence	expected-minus-observed under mandate	trajectory	A
Drift	monotone trend of a statistic	trajectory	A at coarse r
Synchrony	cross-actor correlation	window	O below Δ
Divergence	disagreement between redundant views	event pair	A if both sealers batch independently

3.3 The null-source trichotomy

Every N is one of:

- **Learned**: estimated from F_t. Contaminable — if a fraction π_0 of history is rogue, N is biased toward admitting it. (Conjectured bound: detection power against a persistent actor falls at least linearly in π_0 for learned nulls; needs a specific estimator to prove.)
- **Authored**: given by mandate, inventory, separation rules, redundancy. Incontaminable, incomplete. Completeness is A4.
- **Randomised**: uniform sampling. Needs no N at all; its only parameter is σ.

The trichotomy is exhaustive for detectors whose inputs are F_t, mandate, and σ, since those are the only inputs available. (Established relative to that input set.)

3.4 Composition

Detectors compose by: - **Sequencing**: findings of D_1 set U for D_2 (triage → directed audit). - **Mixture**: at each unit, choose D from a distribution derived from σ. - **Conjunction/disjunction** on findings.

A **configuration** c is a finite composed detector. The **support** C is the set of configurations the deployment may use.

4. Adversary model

4.1 Maturity levels

An adversary is a policy π generating events, conditioned on what it can observe.

- **M0** non-adaptive: π fixed.
- **M1** response-adaptive: observes released findings.
- **M2** schedule-adaptive: observes timing of defender actions.
- **M3** model-adaptive: performs system identification, inferring the active configuration from the response process. Assumed to know C (this document is public).
- **M4** custody-adaptive: can influence σ or the sealer. Outside the framework; violates A2/A3.

4.2 Action classes

A malfeasance is a pattern of one basis type. Action class = pattern type. A campaign is a sequence of classes over time (the lifecycle).

4.3 The evasion map

For each type X there is an evasion transforming a pattern of type X into type e(X) at some cost. (Established as an enumeration; the specific edges are design claims.)

- Cluster → Serial (disperse)
- Serial → Cluster (concentrate)
- Sequence → Relational (split across actors)
- Relational → Sequence (collapse to one actor)
- Drift → step (a single change; caught by Sequence/Absence if the step is typed)

- Synchrony → Serial with lag (desynchronise)
- Absence → performative activity (exercise permissions vacuously; becomes Serial or Cluster)
- Divergence → alignment of views (requires corrupting both sealers; violates A3)

Coverage condition. A configuration c covers a set of types X(c). The deployment is *closed* if for every X in the union of X(c) over c ∈ C, e(X) is also in that union. Closure means no single evasion exits the detected set. (Established as a definition; whether a given C is closed is checkable by inspection of the map.)

4.4 The residual

Iterate the evasion map to its fixed points. One remains: an act that is within mandate, isolated (no relational structure), unordered (no sequence), rare (no cluster, no serial signal), uncorrelated (no synchrony), and singly attested (no divergence). Every basis detector returns null on it by construction. This residual is exactly the set of acts the authored null admits, i.e. mandate error. (Established.) The framework therefore reduces “what behavioural detection cannot see” to a governance object, which is where it should be.

5. Unpredictable deployment

This is the part the construct needs and the part most easily got wrong.

5.1 Why a pure deployment fails against M2–M3

If the deployment uses a single configuration c, an M3 adversary identifies c from the response process and plays a best response π*(c) that either stays in an uncovered type or exploits c’s sampling unit and timing. Against any fixed c with an uncovered type or a learnable schedule, detection probability can be driven toward zero. (Established: best-response argument.)

5.2 Mixed deployment

Let μ be a distribution over C, with the realised configuration at each unit drawn from μ using randomness derived from σ. Define detection probability d(c, π). The defender’s guarantee is

$$v = \max_{\mu} \min_{\pi} E_{c \sim \mu}[d(c, \pi)]$$

and by minimax for finite games the adversary’s best response against μ cannot push detection below v. (Established for finite C and finite π-space; the extension to adaptive π is a repeated game and the bound holds per period under A2.)

Design consequence. Closure of C is not enough; μ must put weight on every type in every period, or the adversary waits for the period in which the type it needs is unweighted. Uniform-ish μ over basis types is the safe default until a specific rule set allows optimisation.

5.3 Independent conspiracy

Each structural component (observer scheduler, sampling-unit selector, batch jitter, release scheduler, configuration mixer) draws its randomness from a distinct stream derived from σ: $\sigma_k = \mathrm{KDF}(\sigma, k)$. Properties (established):

- **Independence in observation**: knowing one component's realised behaviour reveals nothing about another's, since the streams are computationally independent.
- **Joint reproducibility in verification**: at reveal, every component's behaviour is recomputable from σ, so the whole deployment is auditable as one object.

That is the precise form of "independently conspire": independent to the adversary, jointly determined to the auditor.

5.4 Response indistinguishability

Let H be the adversary-observable history: timing of audits, reveals, releases, observer toggles, batch boundaries, and any released findings. The unpredictability requirement is

$$P(H \mid c) / P(H \mid c') \leq \exp(\varepsilon) \text{ for all } c, c' \in C \text{ and all } H,$$

so the adversary's posterior over the active configuration moves at most a factor $\exp(\varepsilon)$ from its prior. This is the same form as differential privacy with configurations in the role of adjacent datasets. (Established as a definition; the construction below achieves it exactly, $\varepsilon = 0$, for the timing component, and approximately for content.)

Construction: - **Timing.** Draw every defender-visible event time from a homogeneous Poisson process with rate fixed in advance and independent of c. A homogeneous Poisson process is the maximum-entropy point process for a given rate: it carries no information beyond its rate. (Established.) This generalises the exponential toggle intervals already in the construct to *all* defender-visible timing: audits, releases, reveals, batch boundaries. - **Existence of releases.** Release findings in fixed-size batches at Poisson times regardless of how many findings exist; pad with sealed-but-empty slots. The public sees a constant-rate process. Slot contents are visible only under the audit mandate. - **Content of releases.** Cannot be made fully independent of c without withholding findings entirely. Trade-off: delay findings by a Poisson holding time before release, which decorrelates release from trigger. The residual leakage $\varepsilon_{\text{content}}$ is the price of ever publishing anything. (Open: quantify $\varepsilon_{\text{content}}$ for a given rule set.)

5.5 Coverage under mixture is not free

Mixing spends the sampling budget across detectors. Against an M0 adversary a pure best configuration outperforms any mixture; the mixture's value is only against M2+. The construct should treat μ's spread as insurance priced against the assumed maturity of the adversary population. Allocating budget across detectors over time is a bandit problem with adversarial rewards; standard adversarial-bandit guarantees apply if rewards (detections) are eventually observable. (Established in outline; Open: the practical regret at governance-scale budgets.)

6. Recursion and the base of trust

Detector outputs are events in E. A captured detector is an actor. Apply Section 4 to it: it shows as Absence (things it should flag and doesn't), Drift (its flag rate changing), or Divergence (disagreement with a redundant detector). Detectors can therefore be monitored by detectors.

The recursion terminates because the randomised null has no learnable structure and no null to corrupt. Its only parameter is σ. So the whole tower rests on two externally supplied objects:

1. a monotone external clock (A1), and
2. an unpredictable committed seed (A2),

plus the two conditions that nothing writes past the sealer (A3) and that the mandate is right (A4). Everything else is derived. The mathematics does not remove trust; it concentrates it into four named places so custody can be designed for them.

7. What is established, conjectured, open

Established (by composition of standard results) - E1 The substrate is a filtration; detectors are adapted processes; the past is annotatable, not editable. - E2 Commit-reveal inherits its unpredictability from the external clock alone. - E3 Δ is an evidentiary parameter; attestation grade is a property of every statistic. - E4 The null-source trichotomy is exhaustive for the given inputs. - E5 The residual of the evasion map equals mandate error. - E6 Pure deployments are beaten by M3 adversaries; mixed deployments have a minimax floor v per period. - E7 Derived streams give independence-in-observation with joint reproducibility. - E8 Homogeneous Poisson timing is maximum-entropy and yields ε = 0 for the timing component of response indistinguishability.

Conjectured - C1 Learned-null power degrades at least linearly in the contaminated fraction π_0. - C2 The eight basis types span the patterns definable over $(E, \prec, \rho)$ with first-order statistics on labels; higher-order or content-dependent patterns may lie outside the basis. - C3 For governance-scale substrates, a near-uniform μ over basis types is within a small constant of the optimal μ for any closed C.

Open - O1 Compute v for a concrete rule set and adversary class; it is a number, not a principle. - O2 Quantify $\varepsilon_{content}$ for a chosen release policy. - O3 The bandit allocation's practical regret at small budgets. - O4 Whether C2 holds; if not, extend the basis. - O5 The recognition rate q (v0.2, Open 18) enters $d(c, \pi)$ multiplicatively for content-judged units and is unmodelled here.

8. What a monitor for any codebase must expose

Setting-agnostic minimum interface, so the framework applies to any substrate:

1. An event hook emitting (actor, type, target, hash, references) for every action, including reads and detector findings.
2. A batching layer with authority-attested T, recording Δ and gap flags, with boundaries drawn from σ_{batch}.
3. A seed service: commit, derive per-component streams, reveal; custody split per the construct.
4. A response scheduler that owns *all* defender-visible timing, Poissonised at a fixed rate.
5. An attestation-grade tag on every statistic, computed from its required resolution against Δ.
6. A configuration mixer drawing from μ per sampling unit, logging the draw as a sealed event.

None of these knows what the records are about.

▸DISCLAIMER

Cite this version

```
Addendum M — Mathematical Strategies for Timelined Detection, version 0.1 (September 2026). agenticgovernance.digital/downloads/addendum-m-v0-1.pdf
```

Licensed CC BY 4.0. Reuse is free, including commercially, provided the author is credited and the reuse does not suggest that the author endorses it. This link is frozen at version 0.1; /addendum-m always shows the current one.