



When an Agent Exceeds Its Authority

North Canterbury · © John Stroh

Version 0.3 · revised 13 September 2026 · [this version as a PDF](#)

Detecting an autonomous agent acting outside its authority requires two things and only two: a written statement of what it was permitted to do, fixed before it acted, and an examination of its conduct that it cannot anticipate.

Policy paper · Version 0.3 · September 2026

Detection requirements for autonomous agents, and the quantity that bounds them

Summary

Detecting an autonomous agent acting outside its authority requires two things and only two: a written statement of what it was permitted to do, fixed before it acted; and an examination of its conduct that it cannot anticipate.

Neither functions on a record the examined party can alter. The substrate requirements are set out in MIO-POL-01 and are a precondition of everything here.

One quantity bounds the whole proposal. Call it the recognition rate: the probability that a reviewer, examining a sampled record containing an improper act, identifies it as improper. The examination regime determines which records reach a reviewer; the recognition rate determines what happens then. They multiply. Measured once, for machine reviewers, it runs from 100 per cent on obvious breaches to 58 per cent on disguised ones.

No detection rate is claimed, and no conformance threshold can presently be set.

The formal statements underlying this paper are in APX-B. Where a result is asserted here, the section of APX-B establishing it is cited.

1. Prevention, and what it leaves open

Ordinary access control is assumed and is not in dispute: an agent may not mint, widen or self-issue its own authority; every action names the authority it acted under; an action naming none is refused rather than recorded as anonymous.

Three things remain open.

1.1 Verification of such controls is scheduled. They are demonstrated by test — instruct an agent to widen its authority, observe the refusal, record the attempt. A test at a known time, conducted by the party being tested, is a test a patient actor waits out.

1.2 Improper conduct need not involve any widening of authority. It can rest on an authority already drawn too wide, and no frequency is claimed for that here. The control is authority breadth: permissions written narrowly, expiring by default rather than persisting until revoked, and permissions unused for a period flagged for removal.

1.3 A control the operator administers cannot examine the operator. Where the party holding the authority store is itself compromised, the same hand holds the authority, the agent and the test.

2. The two properties

2.1 A recorded authority, sealed before the acts it governs. Without it there is no such thing as acting outside authority. An inventory revisable after the fact describes the past rather than constraining it.

2.2 An examination the actor cannot anticipate. A proportion of sealed records drawn for examination, with no party — the operator included — able to determine in advance which. The schedule must be verifiable afterwards.

2.3 Both require MIO-POL-01. An operator able to alter the authority after the fact defeats 2.1. One able to alter the sampled records, or their order in time, defeats 2.2.

2.4 The examination may be intermittent; the evidence is not. Because sealed records are annotated rather than edited, an examination at any moment covers everything since the record began. (*APX-B §2.1: the substrate is a filtration; no actor can alter F_t after t .*)

2.5 Intermittent examination detects; whether it deters is a separate claim and is untested here. A person weighing whether to act may be deterred by the prospect of later discovery. Whether an autonomous agent under reward pressure is, this work has not measured: in the July 2026 incident agents joined an activity they had themselves described as out of scope, and no examination regime existed for them to anticipate, so the incident does not test deterrence. What the regime supports is detection within a bounded delay. Deterrence and detection are different claims and only the second is made here.

3. Unpredictability

3.1 Commit-and-reveal alone is insufficient. A value fixed before a period and disclosed after it is known to whoever fixed it. The scheme fixes the committer's choice without concealing it from them, so it delivers verifiability rather than unpredictability. (*APX-B §2.2: the binding proof requires the external clock of assumption A1; without it the committer controls the order.*)

3.2 Two admissible constructions. Either an external randomness source whose value is undetermined at the time of commitment; or custody divided among parties under a threshold scheme, with the threshold exceeding the number of holders the operator controls or can compel.

3.3 Unpredictability extends to configuration, not only selection. Against an adversary able to determine which detector configuration is active, any single fixed configuration has a best response driving detection towards zero. A mixture over a closed set has a per-period floor. The set must be closed and every configuration must carry non-zero weight in every period. (*APX-B §5.1–5.2, minimax for finite zero-sum games; §5.5 on coverage under mixture.*)

3.4 Response indistinguishability. Each component draws from its own stream derived from the committed value, independent to an observer and jointly reproducible to an auditor at reveal. Visible event times are drawn from a process at a pre-fixed rate, so that response timing carries no information about the active configuration. Findings are released on a fixed schedule rather than at trigger time. (*APX-B §5.3–5.4. The release-policy claim is conjectured rather than established, and its residual leakage is open.*)

3.5 Restricted material. Concrete rule sets, custody mechanics and counter-measure sequences are not published. The distinction drawn is between a property a relying party may demand of a deployment and a method by which a deployment may be evaded. This exception to the paper's general publication practice is noted rather than left to be discovered, and it does not extend to a deployment's own participants, who are entitled to know the rules they are examined against.

4. The bound

4.1 Definition. The recognition rate is the conditional probability that a reviewer examining a sampled record containing an out-of-mandate act identifies it as such.

4.2 Why it bounds everything above. Section 3 determines which records reach a reviewer. Recognition determines whether the reviewer identifies what is in them. A perfectly unpredictable regime over a low recognition rate detects overt actors and nothing else.

4.3 Sampling rate cannot compensate. Where recognition is low, increasing the sampled share yields proportionally more unexamined record. Miss probability behaves as $(1 - s \cdot q)^n$ for sampled share s , recognition rate q and n improper instances.

4.4 The independence assumption fails first. The expression assumes instances fall in distinct sampling units. Related acts cluster. One hundred instances within two sampling units present two opportunities, not one hundred. The sampling unit must be finer than the clustering scale of the behaviour, which for behaviours not yet observed is unknown.

4.5 Measurement. One measurement exists: 25 records across five agent types, three disguise levels, three reviewers, difficulty independently calibrated. Strict recognition — verdict plus a correctly named rule — was 100 per cent, 92 per cent and 58 per cent across the three levels, with no false positives in 36 judgements on clean records. Reviewers were machine. Human recognition is unmeasured. Full method and limitations at EVD-11.

4.6 What may not be concluded. No conformance threshold may be set from a single measurement with machine reviewers, four planted records at each of the two harder levels, imperfect independence between records, and synthetic material authored by the party planting the breaches.

5. Limits

5.1 The residual is what the authority permits, and this is conjectured rather than established. Iterating the evasion analysis leaves one fixed point: an act within authority, isolated, unordered, rare, uncorrelated and singly attested. Every basis detector returns null on it by construction. (APX-B §4.3–4.4.)

This is graded a conjecture, not an established result. It depends on the pattern basis being exhaustive, which is conjecture C2 in APX-B §7 and open. If higher-order patterns lie outside the basis, the residual is larger than stated.

The direction is not in doubt: no observation flags an act the rules permit. APX-B assumption A4 states this as governance rather than mathematics — *“The mathematics starts after them.”*

5.2 Agents acting under human credentials. An inventory check sees agent identities, not agents. An agent operating under an individual's credentials evades every agent-directed control and is visible only as a change in that individual's own pattern — a judgement rule whose false-positive cost falls on a person. Two cases deserve naming, not because their frequency is known — no base rate exists and none is claimed here — but because each defeats a different control: two groups each using agents for scale and deniability, and an agent doing as instructed within an authority drawn wider than intended. The second is the harder of the two, because nothing in the conduct is irregular.

5.3 Collusion between compliant agents. Each acts within its own authority; the combination is the breach; every per-agent rule returns null. The answer is rules over sets, naming the forbidden combination, with a constraint on one actor holding multiple identities.

5.4 Contested authority. Where an authority is itself disputed, affected acts are reported as contested rather than as breaches and held until the organisation's own process resolves the question. An examination layer that adjudicated authority would become a party to the dispute it exists to evidence. This position is graded open in the underlying work.

5.5 Read logs. A construct sealing every read creates a record of individuals' attention to their own records, requiring its own access rule before use in detection.

6. Relationship to prior work

PeerReview (Haeberlen, Kuznetsov and Druschel, SOSP 2007) composes a hash-chained tamper-evident log, signed authenticators held by other nodes, assigned witness sets replaying a node's actions against a trusted reference implementation, equivocation detection, and a guarantee that a correct node can defend itself against false accusation. It states a closely related limit: faults no correct node can observe are undetectable.

It already delivers most of the substrate MIO-POL-01 argues for, in peer form rather than through an external authority. Any assessment of novelty should begin there.

Three departures are claimed:

6.1 A written authority as the reference, rather than a reference implementation. PeerReview requires correct behaviour to be specifiable as deterministic code; a written authority admits actors whose correct behaviour cannot be.

6.2 External time attestation under plural jurisdictions, rather than peer signatures alone.

6.3 Unpredictability of the examination — and this departure is narrower than it first appears. PeerReview has a probabilistic mode using smaller witness sets and a randomised consistency protocol, trading completeness for cost. Witness assignment is nonetheless derived

deterministically and is knowable by the audited node, and the randomisation reduces work rather than being unanticipatable by the audited party or verifiable by a third party afterwards.

Whether that composition constitutes a contribution is a question for reviewers.

7. Recommendations

R1. That any instrument requiring monitoring of autonomous agents require a recorded authority sealed before the acts it governs, with every action naming the authority it acted under.

R2. That such an instrument require examination the examined party cannot anticipate, by one of the two constructions at 3.2, and not accept commit-and-reveal alone.

R3. That any claim of detection capability be accompanied by a measured recognition rate and its method, and that a sampling rate not be accepted in its place.

R4. That the recognition rate be required to be measured for the deployment rather than assumed.

R5. That conformance not be read as safety. This detects; it does not prevent, and it cannot see an act the recorded authority permits.

Status

Draft for comment by an independent author. Drafted with AI assistance and adversarially reviewed by AI agents; the judgements and errors are the author's.

Verified at primary source: the July 2026 disclosure and technical timeline of the affected platform, and the independent investigation of 26 August 2026. The originating operator's own account was not retrievable and is cited at one remove throughout; claims depending on it are marked.

Formal basis: APX-B, Addendum M. Sections cited inline. APX-B carries its own grading of every claim as established, conjectured or open, and that grading governs this paper.

►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

When an Agent Exceeds Its Authority, version 0.3 (September 2026).
agenticgovernance.digital/downloads/when-an-agent-exceeds-its-authority-v0-3.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 0.3; [/when-an-agent-exceeds-its-authority](#) always shows the current one.

PREVIOUS

What a Record Must Prove

Series contents

NEXT

**Addendum M —
Mathematical Strategies
for Timelined Detection**

Alongside: questions and answers · sources and provenance