

When 10,000 Agents Act

Authority, Liability and Evidence in the Synthetic Economy

Population-scale delegation and the recognition of autonomous economic action

Conceptual framework developed by Steven Alber

Date of first formulation of the programme: 30 August 2026

This version: **v0.1 — 12 October 2026**

Status: public research note, open for criticism

Programme lineage: Founding Thesis v0.1 → red-team audit → Founding Thesis v0.2 → hostile verification → WP-001 v0.1 → hostile verification Round II → WP-001 v0.2 → External Validation → WP-001 v0.3-RC1 → **RN-001 v0.1**

Correspondence and criticism: research@thesyntheticconomy.org

This note has not been peer-reviewed. No external scholarly reviewer has read it. The adversarial critique in Section 13 is internal critique produced within the programme and must not be described as independent external review.

Prepared with AI-assisted research, adversarial analysis and primary-source verification. Selection of propositions, the decision as to what may be claimed, and final responsibility are the author's.

This note does not replace or revise Working Paper No. 1 v0.3-RC1, which remains a private review candidate. Where this note bears on a Working Paper proposition, the bearing is recorded as an open issue for a controlled revision in Appendix C.

Abstract

An enterprise that deploys ten thousand autonomous economic agents must answer four questions about every one of their actions: who authorised it, who bears the consequences, what legal standing the mandate has outside the system that issued it, and what evidence would let the action be contested afterwards. This note argues that the first three questions are being answered — unevenly, but visibly — by a commercial agent-governance layer that reached general availability during 2026, while the fourth is not being answered by the same layer, and that the asymmetry between them becomes constitutive rather than merely inconvenient at population scale.

The argument rests on documented absence as much as on documented presence. Primary vendor documentation for enterprise agent identity governance specifies, in detail, how an agent acquires a distinct identity, how its permissions are scoped through templated blueprints, which human sponsor is accountable for it, which approver must clear its access requests, and when that access lapses. The same documentation specifies sign-in and authentication logging. It does not specify a retained record that links a particular executed action to the particular mandate, scope and approval under which it was taken, across delegation hops — what this note calls **mandate provenance**. Independently and from inside the standards process, an Internet-Draft states that the governing token-exchange standard permits nested actor claims but defines no interoperable rules for preserving, extending, disclosing and validating a delegation path across successive exchanges, leaving chain continuity and audit evidence to private convention; that draft is an individual submission and has not been adopted by a working group.

Three propositions follow. First, the authorisation layer and the evidence layer are being built at different rates and by different actors, and only the first is converging. Second, this asymmetry is tolerable at small numbers and corrosive at large ones, because the cost of reconstructing an authorisation after the fact rises with the number of concurrent, short-lived, interacting mandates while the expected value of any single reconstruction falls. Third, neither the international instrument track nor the converging private standards body is currently building the missing layer: the UNCITRAL working group that produced the Model Law on Automated Contracting has completed that work and moved to other subjects, and the standards body now consolidating the commercial agent-payment stack has published no specification and makes no claim about legal recognition.

The note gives full weight to the strongest objection: that absence from a documentation page is not absence from a product, that enterprise platforms carry extensive audit machinery not described on the pages consulted, and that evidence law already allocates the risk of unproducible records through burdens of proof, business-records doctrine and adverse inference — so that what looks like an institutional gap may be an ordinary allocation of litigation risk requiring no new institution at all. That objection is not defeated here. It is made falsifiable, and the conditions under which this note's propositions should be abandoned are stated in Section 15.

0. Position of this note in the programme

The Synthetic Economy is an independent research programme on the political economy of delegated non-human agency. Its canonical historical baseline is the Founding Thesis v0.2 (30 August 2026), preserved unchanged and superseded in substance. Its current scientific paper is Working Paper No. 1 v0.3-RC1 (31 August 2026), a private review candidate prepared for external scholarly criticism, which defends two propositions concerning population effects on dyadic doctrine and the boundary between private technical authorisation and sovereign legal recognition. The programme's method and its complete record of retired and reformulated claims are held in the Revision & Evidence Ledger v0.3.

This note does not replace Working Paper No. 1 and does not revise it. WP-001 remains the programme's canonical paper on the recognition boundary. RN-001 takes a different cut through the same territory: where WP-001

asks what the boundary between private and sovereign recognition implies for doctrine and rents, RN-001 asks what happens inside an organisation that has already crossed into operating agents at population scale, and what can be proved about their actions afterwards.

Where RN-001 bears on a WP-001 proposition it says so explicitly and records the bearing as an open issue for a future controlled revision of WP-001 rather than altering WP-001's text. Two such bearings are identified, at Sections 4 and 9, and are carried into the Ledger supplement as entries S-05 and S-06. Neither is treated here as settled.

RN-001 is published rather than circulated privately because its propositions are more exposed than WP-001's and benefit from early attack. It is a research note, not a working paper: shorter, narrower, and explicitly provisional.

1. The claim

Three propositions are defended. Each carries a claim-class tag in the programme's notation, where **[FACT]** marks what a primary source states, **[DOCUMENTED PRACTICE]** marks what a primary source shows a named actor doing, **[INFERENCE]** marks what follows from the convergence of such facts, **[HYPOTHESIS]** marks what requires empirical or institutional testing, and **[NORMATIVE]** marks a proposal about what ought to be done.

Proposition RN-1 — the authorisation–evidence asymmetry

[INFERENCE] In the enterprise agent-governance layer that reached general availability during 2026, the controls that *confer* authority on an autonomous actor are specified in substantially greater detail than any retained record that would allow the exercise of that authority to be *reconstructed and contested* afterwards. Authentication logging is documented. Mandate provenance — the reconstructable lineage linking a specific executed action to the specific mandate, scope, blueprint inheritance and human approval under which it was taken, across delegation hops, together with the validity state of that mandate at the moment of execution — is not specified in the governance documentation consulted, and the standards work that would make it interoperable across systems is an unadopted individual submission.

Primary support. The identity platform describes agent identity blueprints as templates with parent–child relationships, so that one policy applies to many agents, and names sign-in and audit logging for agent authentication and activity (L1-01). The governance documentation specifies human sponsors "accountable for making decisions about its lifecycle and access", access granted through entitlement-management access packages with requests routed to designated approvers, OAuth 2 delegated scopes inherited from the parent blueprint, three distinct request pathways, expiry-based lapse of access, and automatic transfer of sponsorship to the sponsor's manager on departure (L1-02). The same governance page makes no statement about audit logs or evidence of agent actions, and the release-notes page does not cover audit and logging (L1-02, L1-03). Separately, draft-mw-oauth-actor-chain-01 states that RFC 8693 permits an `act` claim to carry nested prior actors but defines no interoperable rules for preserving, extending, disclosing and validating a delegation path across successive exchanges, leaving implementations to invent their own conventions for chain continuity, disclosure control and audit evidence (L1-04).

What the proposition does not assert. It does not assert that the products concerned lack audit capability. It asserts that the documented governance surface specifies conferral in more detail than provenance, and that the interoperable form of provenance is unadopted. The distinction matters and is pressed hard against this proposition in Section 13.

Proposition RN-2 — scale changes the economics of contestability

[HYPOTHESIS] The authorisation–evidence asymmetry is tolerable at small numbers of agents and corrosive at large ones, not because reconstruction becomes impossible but because its economics invert. As the number of concurrent, short-lived, interacting mandates rises, the cost of reconstructing any particular authorisation rises with the volume of candidate records and the number of hops, while the expected value of reconstructing it falls with the declining share of total harm attributable to any single action. Beyond some threshold, rational parties stop contesting individual actions and contest only aggregates — and aggregates are precisely what dyadic doctrine is least equipped to adjudicate.

This is a hypothesis about cost structure, not a claim about current practice. It is testable: Section 15 states what measurement would support or defeat it, and the research agenda sets out the studies that would produce it.

Proposition RN-3 — neither track is currently building the missing layer

[FACT] The UNCITRAL working group whose work produced the Model Law on Automated Contracting has completed that subject and is now mandated to work on data provision contracts, followed by consolidation of e-commerce texts and paperless trade; its 70th session (March 2026) had data provision contracts as the sole agenda topic (L1-05).

[FACT] The standards body into which the commercial agent-payment stack is converging formed its Agentic Authentication Technical Working Group on 28 April 2026, received the Agent Payments Protocol from Google and the Verifiable Intent framework from Mastercard, has published no specification, and says nothing in its announcement about legal recognition or regulation (L1-06).

[INFERENCE] The layer that would make mandate provenance both interoperable and legally cognisable is therefore not currently the active work programme of either the international instrument track or the converging private standards track. The one place it is being worked on in public is an individual Internet-Draft that no working group has adopted (L1-04).

2. Research scope, and what this note does not claim

This note concerns delegated economic action inside and across organisations operating autonomous agents at a scale where individual supervision is not performed. "Ten thousand" is a scale marker, not a measured population: no real deployment of that size has been studied here, and SE-001 is explicitly a conceptual instrument.

For the avoidance of doubt, and in application of the programme's revision method, RN-001 does **not** claim:

- that agent identity, scoped delegation, approval workflows or lifecycle governance are novel as computer science or as enterprise practice — they are longstanding, and the documentation cited shows them being extended to a new class of principal;
- that machine-to-machine delegation is legally novel merely because the delegate is a machine; the programme retired that claim in Ledger entry 2 and does not revive it;
- that the vendors discussed lack audit logging, or that any named product is deficient;
- that any specific market size, transaction volume, adoption rate or revenue figure obtains in agentic commerce; no such figure was verified and none is asserted;
- that the September 2026 Copilot announcement has any particular content; its primary text could not be retrieved in this cycle and nothing here depends on it (L3-01);

- that private ordering has displaced, or is in the process of displacing, sovereign legal recognition; and
- that evidence law is inadequate. Section 13 treats the sufficiency of existing evidence doctrine as the strongest objection to this note and does not defeat it.

Microsoft documentation is used as the best available primary window onto how the enterprise agent-governance layer is actually specified. It is an illustrative case. The programme's object is the institutional layer, not any vendor.

3. Population-scale delegated agency

Principal–agent theory and the law of agency were both built around a relationship: one principal, one agent, one grant of authority, one act whose attribution is disputed. The economic literature studies the relationship's incentive problem; the legal doctrine studies its attribution problem. Both scale by repetition — a firm with a thousand employees is analysed as a thousand relationships, with organisational structure as an overlay.

The question this note asks is whether that repetition remains adequate when the agents are machine actors created from templates. Three features of the documented architecture bear on it, and each cuts partly against the idea that anything fundamental changes.

Templated authority. Agent identity blueprints act as templates with parent–child relationships so that a single policy applies to many agents, and new agent identities inherit OAuth 2 delegated scopes from the parent blueprint (L1-01, L1-02). **[INFERENCE]** This is organisationally familiar — it resembles a role or a job description — but it has a property that a job description lacks: the inheritance is machine-evaluated at authorisation time, so a change at the blueprint level propagates to the whole population without any per-agent act. The unit at which authority is actually set is the blueprint, while the unit at which action is attributed remains the agent. Analytically, the relationship is no longer the natural unit; the population under a blueprint is.

Accountable humans, but not necessarily. Governance documentation identifies sponsors as "human users accountable for making decisions about its lifecycle and access" and calls the assignment of a delegated human one of the most important parts of governing agent identities, while not stating that a sponsor is mandatory or enforced; expiry notifications reach the sponsor only if one is set (L1-02). **[INFERENCE]** The architecture therefore anticipates, without excluding, the orphaned agent: an actor with live permissions and no accountable human. Whether orphaning occurs in practice is an empirical question, and the agenda proposes the study that would answer it. It is worth noting that the documented mitigation — automatic transfer of sponsorship to the departing sponsor's manager — addresses departure, which is the easy case, and not abandonment, which is the hard one.

Two kinds of authority, separated by design. The platform distinguishes an "On behalf of agent access policy" for agents acting on behalf of users from an "Autonomous agent access policy" for agents without user context (L1-03). **[FACT]** This is a significant piece of architecture for the programme's purposes, because it is the vendor's own recognition that delegated action and autonomous action require different authorisation treatment. **[INFERENCE]** It also means that the question "on whose behalf did this agent act?" has, at the level of policy, a machine-readable answer — which is a point in favour of attribution being easier than the programme has sometimes assumed, and is recorded as such.

What this does not establish. None of this shows that agency doctrine fails. The Model Law on Automated Contracting attributes the output of an automated system to the person on whose behalf it is used (L2-01), and that rule is indifferent to how many systems there are. The honest summary is that population scale changes the *unit of control* without, on current evidence, changing the *unit of attribution* — and that the interesting problems live in the gap between those two units, not in the failure of either.

4. Recognition and jurisdiction

The programme has previously distinguished private technical authorisation from sovereign legal recognition, and WP-001 defends a narrowed form of that distinction. RN-001 adds a distinction *within* the private side, because the documented architecture now supports it.

Four states should be kept apart:

1. **Technical authorisation** — the authorisation decision succeeded; the action executed.
2. **Attributable authority** — the action is attributable to a legal principal under a rule such as the Model Law's attribution provision.
3. **Contractual recognition** — a counterparty has agreed, by contract or scheme rule, to treat mandates of this form as binding.
4. **Sovereign legal recognition** — a legal order gives the mandate effect as such, including where it was constituted abroad.

[INFERENCE] These four come apart in both directions, and conflating them is the standing error in this area. An action may execute technically and bind nobody; it may bind a principal despite a technical authorisation failure, because attribution does not depend on the integrity of the authorisation system; and it may be contractually recognised across borders while having no sovereign recognition anywhere. Scheme rules and network participation agreements do a great deal of work that is easily mistaken for law.

The point for population scale is that states 1 and 3 scale automatically — they are machine-evaluated and contractually pre-agreed — while states 2 and 4 are resolved by human institutions, one dispute at a time.

[HYPOTHESIS] The divergence in scaling behaviour, not the divergence in substance, is what makes the boundary economically significant.

Bearing on WP-001. WP-001's Proposition 2 holds that sovereign legal recognition of foreign-constituted machine mandates remains substantially unbuilt. The UNCITRAL finding (L1-05) bears on this: the international track that might have built it has concluded its AI-and-automation work and moved on. On its face this strengthens Proposition 2 — the gap is not merely unbuilt but currently unattended at the instrument level. It also weakens any version of the proposition that assumed an active work programme would close the gap on a five-year horizon, which is relevant to WP-001's stated falsifier. **This bearing is recorded as an open issue for a future controlled revision of WP-001 and is not applied to WP-001 here.** See Ledger supplement entry S-05.

5. Recursive delegation and the evidentiary residue

The programme retired, in Ledger entry 2, the claim that recursive delegation is novel as computer science; certificate path constraints, capability systems and attenuable tokens have implemented it for years. In Ledger entry 15 it reframed what survives: what the protocol layer externalises at hop N is not a legal problem of attribution but an **evidentiary residue**. RN-001 can now state that residue with a named primary source rather than as an inference.

[FACT] draft-mw-oauth-actor-chain-01 states that RFC 8693 permits an `act` claim to carry nested prior actors, but does not define interoperable rules for preserving, extending, disclosing and validating a delegation path across successive exchanges, with the result that implementations invent their own conventions for chain continuity, disclosure control and audit evidence (L1-04).

[DOCUMENTED PRACTICE] The draft's response is instructive for the programme's purposes because of how it decomposes the problem. It defines six profiles from two **evidence models** — *declared* and *verified* — and three

disclosure modes — *full, subset and actor-only* — and, for verified profiles, actor-signed step proofs with cumulative commitment state and a stable workflow identifier (L1-04).

[INFERENCE] Three observations follow, and they are the analytical core of this note.

First, evidence is now explicitly a design parameter rather than a by-product. The distinction between a declared chain and a verified one is the distinction between a record that asserts what happened and a record that can demonstrate it. That these are offered as alternative profiles means the strength of the evidentiary residue is a configuration choice made by the deploying party — which is to say, by the party most likely to be the defendant.

Second, disclosure is a separate axis from evidence. A chain may be fully verified and disclosed only as "actor-only". The capacity to prove and the obligation to reveal are orthogonal, and a system can be designed to be provable to its operator and opaque to a counterparty. This is the precise technical shape of a contestability problem.

Third, and most important for the programme's claim about the sovereign boundary: this work is unadopted. It is an individual submission on the Standards Track, not endorsed by the IETF, with no working group behind it (L1-04). The gap has been identified from inside the standards process by authors at Oracle, JPMorgan Chase, Telefónica and Aryaka — which is evidence that practitioners in payments and infrastructure regard it as real — and the response to it has not been institutionalised.

Counter-consideration, given its full weight. An unadopted draft is weak evidence of anything except that four engineers thought the problem worth writing up. Many Internet-Drafts expire unread. The draft's own statement that implementations invent their own conventions is simultaneously evidence that the gap exists *and* evidence that it is being filled privately, case by case, which may be perfectly adequate. The honest reading is that the draft establishes the absence of an *interoperable* standard, not the absence of working practice, and RN-001 claims only the former.

6. Collective and systemic behaviour

The programme's position on emergent collective behaviour is recorded in WP-001 and is not extended here. In summary and without strengthening: the phenomena are demonstrated in controlled settings; sensitivity to design, robustness and external validity are contested by the field including by the original authors; and the transfer from controlled experimental environments to deployed markets is an exposed inferential step (L2-06). The appellate authority runs to two poles, dismissal where no competitor data was pooled and a sustained pleading where non-public data was pooled through a shared algorithm (L2-04, L2-05).

RN-001 adds one observation specific to population scale within a single enterprise, which is a different setting from the inter-firm coordination cases.

[HYPOTHESIS] Where a population of agents inherits scope from a shared blueprint (L1-01, L1-02), their behaviour is correlated by construction rather than by learning. A blueprint change is a simultaneous change to the effective policy of every agent beneath it. This is a mechanism for correlated action that requires no emergence, no communication and no learning dynamics at all — and it is therefore both more mundane and more tractable than the phenomena in the experimental literature. It is also, for the same reason, a more plausible near-term source of aggregate harm: a misconfigured blueprint is an ordinary operational failure with population-wide reach.

[INFERENCE] If that is right, the first serious population-scale incident is more likely to resemble a configuration cascade than an emergent conspiracy, and the institutional response appropriate to it is closer to change management and audit than to competition law. The programme should not assume its most interesting hypothesis is also its most probable scenario. This is deliberately stated as a point against the programme's own emphasis.

7. Liability and attribution

Attribution is the part of this field where existing law is strongest, and the note says so.

[FACT] The Model Law on Automated Contracting attributes the output of an automated system to the person on whose behalf the system is used (L2-01). **[DOCUMENTED PRACTICE]** The platform architecture independently distinguishes on-behalf-of from autonomous operation at the level of access policy (L1-03), which means the "on whose behalf" question has a machine-readable answer in at least one major deployment pattern.

[INFERENCE] Attribution to a principal is therefore not, on current evidence, the hard problem. The hard problems are adjacent to it:

- **Scope, not identity.** Agency doctrine turns on whether the act fell within actual or apparent authority (L2-03). Where scope is inherited from a blueprint and evaluated dynamically, the question "was this within scope?" is answerable only against the state of the blueprint at execution time — which is a records question, returning the analysis to Proposition RN-1.
- **Revocation and in-flight action.** Access ends automatically at an assignment's expiry if the sponsor takes no action (L1-02), and the documentation consulted does not state that disabling an agent revokes its access. What happens to commitments already initiated by downstream agents at the moment authority lapses is, on the documents consulted, not specified. Scenario D of SE-001 develops this.
- **Intermediaries.** Where a platform operator evaluates the authorisation and a scheme operator defines the mandate format, both have made a contribution to the action without being its principal. Existing law has several frames for this — intermediary liability, negligent provision of a service, scheme-rule contractual allocation — and no settled choice among them.

[NORMATIVE — flagged as proposal, not finding] Nothing in this note supports creating a new liability subject. The programme's default remains principal-default: consequences attach to the principal through the mandate chain. Proposals to attach liability to the agent itself, or to create patrimonial shells, are recorded in the programme's earlier documents as institutions to be studied, not recommended.

8. Evidence and contestability

This is the section the note exists for.

To contest an autonomous action after the fact, a party must be able to establish at least five things: that the action occurred and in what terms; which agent identity took it; which mandate purported to authorise it; what the scope of that mandate was *at the moment of execution*, including everything inherited from its blueprint; and who approved that mandate and when. Call the retained record sufficient to establish all five **mandate provenance**.

[FACT] The documented governance surface specifies the inputs to all five. Agent identity is a first-class construct (L1-01). Mandates are granted through access packages with named approvers (L1-02). Scope is inherited from blueprints (L1-02). Validity has an expiry (L1-02). Sign-in and audit logging for agent authentication and activity is named (L1-01).

[FACT] What the consulted pages do not state is that these are retained *as a linked record*. The governance overview makes no statement about audit logs or evidence of agent actions; the release-notes page does not cover audit and logging; the only related statement is that a portal displays an agent's access, activity and lifecycle information, without describing what is logged or retained (L1-02, L1-03).

[**INFERENCE**] The gap is not between logging and no logging. It is between **event logging** and **provenance**. An authentication log answers "did this identity authenticate, when, from where?" Provenance answers "under which grant, of what scope, approved by whom, still valid at the time?" The first is a security artefact and is clearly present. The second is a legal artefact, and on the documents consulted it is not specified.

[**INFERENCE**] The standards evidence converges on the same distinction from the other direction. That draft-mw-oauth-actor-chain-01 must *introduce* evidence models and actor-signed step proofs, and must observe that implementations currently invent their own conventions for audit evidence (L1-04), is an independent indication that provenance across hops is not something the existing stack provides.

[**HYPOTHESIS**] **The contestability threshold.** Combining this with Proposition RN-2: there exists, for any given dispute-resolution forum, a population size and mandate-churn rate above which reconstructing mandate provenance for a specific action costs more than the action's individually attributable harm. Above that threshold, individual contestation is economically irrational even where it is legally available, and the practical consequence is that only aggregates get litigated. The threshold is a function of record design, not of agent count alone — which is why it is a research object rather than a lament.

Why this may be a false alarm. Evidence law has handled unproducible records for a very long time, and it handles them by allocating risk rather than by requiring better records. A party that cannot produce the record bears the burden; business-records doctrine determines admissibility; spoliation rules and adverse inference punish destruction; discovery obligations compel production of what exists. On this view, an enterprise that cannot reconstruct mandate provenance simply loses disputes it would otherwise win, and will therefore build provenance for commercial reasons without any institutional intervention. **This is the strongest objection to the whole note.** It is addressed, not dismissed, in Section 13.

9. Permission infrastructure and economic power

Who controls the infrastructure through which agents acquire permission to transact is the programme's standing question about rents, and RN-001 contributes two verified data points without asserting any market outcome.

[**FACT**] Extending the platform's security features to agents requires a specific licence, licensed per user (L1-01), and the platform's release notes state that agent registries are converging under a single product (L1-03). [**FACT**] On the payments side, two major contributors donated their agent-payment and intent-verification frameworks into one standards body, which has published no specification (L1-06).

[**INFERENCE**] Both movements are toward consolidation of the permission layer — in one case commercially, within a vendor's identity estate, and in the other case institutionally, within a standards body. Consolidation of a permission layer is the precondition for permission rents; it is not evidence that rents exist. The programme has already retired the claim that permission-rent economics is theoretically novel, and the analytics remain Coasean rights, delineation costs, two-sided access pricing and capture theory applied to a new object (L2-08).

[**HYPOTHESIS**] The economically interesting variable is not the price of permission but the **cost of switching the authority that recognises you**. An agent population whose identities, blueprints and inherited scopes are expressed in one vendor's constructs has a migration cost that rises with the population and with the depth of blueprint inheritance. If that is right, the rent accrues to lock-in rather than to pricing, and is invisible in any price series — which has an uncomfortable methodological implication for the programme's proposed authority-price index.

Bearing on WP-001. WP-001 develops recognition costs and permission rents as an empirical-methodological programme oriented to price formation. The switching-cost hypothesis suggests that a price-based index may

measure the wrong thing. **Recorded as an open issue for a future controlled revision of WP-001; not applied here.** See Ledger supplement entry S-06.

Avoided. No claim is made about market size, concentration ratios, transaction volumes or revenue. No verified figure was available and none is asserted (L3-03).

10. Institutional responses

Four responses are available in principle. They are set out with their trade-offs; the note recommends none.

Private standards. Fastest, already in motion, and the only track with an actual proposal for provenance (L1-04, L1-06). Weaknesses: adoption is voluntary, the evidence model is a configuration choice left to the party who will be the defendant, and the standards body consolidating the stack makes no claim about legal recognition (L1-06) — so private standardisation does not by itself produce cognisability.

Commercial law and courts. Requires no new institution. Evidence doctrine allocates the risk of missing records; agency doctrine attributes; the Model Law supplies a default attribution rule for automated systems (L2-01). Weaknesses: operates after the fact, one dispute at a time, at human tempo, and reaches only parties before the court. If Proposition RN-2 is right, the forum is reached only when the aggregate is large enough to be worth litigating.

Regulators and supervisors. Can compel record-keeping *ex ante*, which is the direct response to the asymmetry. Weaknesses: sectoral, jurisdictionally bounded and — on the one instrument track examined here — not currently engaged with machine principals at all (L1-05).

Insurance and audit. Price the residual risk and, in doing so, create a private demand for provenance that no regulator imposed. Weaknesses: underwriting a risk requires a loss history that does not yet exist, and an audit opinion on a record system is not evidence of any particular action.

[INFERENCE] The four are not alternatives. The likeliest path is that insurance and audit generate commercial demand for provenance, private standards supply its form, and courts determine *ex post* whether the form supplied is sufficient — with sovereign recognition arriving last, if at all, and only for cross-border cases. That sequence is a conjecture about institutional order, stated as such, and is the subject of research track five.

11. Research Case SE-001 — The 10,000-Agent Enterprise

SE-001 is a structured conceptual instrument, published in full as a separate document and summarised here. It posits a hypothetical enterprise operating ten thousand autonomous economic actors and develops six scenarios: a valid internal mandate whose legal effect a counterparty disputes; a delegation chain extending beyond the principal's intended scope; thousands of individually lawful transactions producing a harmful aggregate; a mandate revoked while dependent agents are still acting; a cross-border action accepted technically but disputed legally; and an automated decision with incomplete evidence of authorisation.

Each scenario states its parties and assumptions, the economic or legal mechanism engaged, at least two rival explanations, the evidence that would discriminate between them, and — required in every case — what the scenario *cannot* establish. No real deployment of this size has been studied. The case generates hypotheses and tests their coherence; it is not evidence for any of them.

12. Research agenda

The five-track agenda *Population-Scale Agency and the Institutions of the Synthetic Economy* is published as a separate document. Its tracks are: delegation and authority; economic coordination and systemic behaviour; legal

recognition and liability; evidence, verification and contestability; and permission infrastructures and institutional power. Each track states its research problem, primary questions, existing theoretical foundations, candidate evidence, testable hypotheses, methodological obstacles, potential empirical studies and specific next outputs.

The agenda is an intention, not an accomplishment. No track has been completed. Several depend on access to deployment data that the programme does not have and has no present route to obtaining; where that is so, the agenda says it.

13. Adversarial critique — internal

What follows is internal critique produced within the programme's own method. **It is not independent external scholarly review and must not be represented as such.** Two separated passes were conducted: the first against the propositions as drafted, the second against the surviving forms after revision.

Pass I — against the propositions as drafted

C1. Absence from a documentation page is not absence from a product. (*Against RN-1.*) Overview pages are written for orientation and defer detail to linked articles; the governance overview explicitly does so. A platform of this scale has extensive audit infrastructure, and the *what-is* page does name sign-in and audit logging for agent authentication and activity. Inferring an evidentiary gap from the silence of two pages is weak method, and the programme has previously had to correct exactly this failure mode.

Response, and the revision it forced. The criticism is substantially correct and RN-1 was narrowed because of it. The original drafting implied that audit capability was absent; it now asserts only that the *documented governance surface* specifies conferral in more detail than provenance, that the pages consulted do not describe what is retained, and that the *interoperable* form is unadopted. The verified negatives in the register are explicitly scoped to named pages at a named date. **Residual exposure: high.** A single vendor document specifying retained mandate-provenance records would defeat RN-1 in its current form, and the author has not read the entire documentation set.

C2. Evidence law already solves this by allocating risk. (*Against RN-1 and RN-2.*) Legal systems do not require that records exist; they decide who loses when they do not. Burden of proof, business-records doctrine, spoliation and adverse inference, and discovery obligations together create a strong commercial incentive to retain provenance without any institutional innovation. An enterprise that cannot prove authorisation simply loses, and will build the records to avoid losing.

Response. This is the strongest objection and it is not defeated. Three qualifications are offered, none decisive. First, risk allocation works where a forum is reached; Proposition RN-2 is precisely the claim that at scale the forum is reached only for aggregates. Second, adverse inference disciplines destruction more reliably than it disciplines never having recorded. Third, the incentive argument predicts that provenance will be built *where the expected litigation cost justifies it*, which is an argument for uneven provision, not universal provision — and unevenness across a cross-border population is itself the research object. **Standing: RN-1 survives as a claim about documentation and standardisation; the claim that this constitutes an institutional gap is downgraded to hypothesis.**

C3. The Internet-Draft proves nothing. (*Against RN-3.*) Individual submissions are cheap. An unadopted draft by four authors is not evidence of an institutional gap; it is evidence that four people wrote a document. The programme has been burned before by treating unadopted drafts as load-bearing — Ledger entries 16 and 23 record exactly this.

Response. Accepted in part. The draft is relied on for two things only: its *description* of what RFC 8693 does and does not define, which is checkable against RFC 8693 itself and is not a matter of the authors' opinion; and its *adoption status*, which is a fact about the IETF process. It is not relied on as evidence that the gap matters, that the

proposed profiles are good, or that anyone will implement them. The affiliations of the authors are noted as weak corroboration that practitioners regard the problem as real, and are given no more weight than that.

C4. "Population scale" is doing rhetorical work. (*Against RN-2.*) Ten thousand agents is a number with no analytical content. Enterprises already operate far more than ten thousand automated processes — every scheduled job, every integration, every RPA bot — and the institutions have coped. The note does not establish that agents differ from scripts in any way that matters to the argument.

Response, and the revision it forced. Largely accepted, and it is the reason Section 3 concludes that scale changes the unit of control rather than the unit of attribution. The distinguishing feature offered is not autonomy or number but **dynamic scope inheritance with machine-time evaluation** — a scheduled job's permissions are static and inspectable; a blueprint-inherited scope evaluated at authorisation time is neither. Whether that distinction carries the weight placed on it is unresolved, and it is listed as a falsification condition.

C5. The note does not prove its own novelty. Records management, evidence preservation and identity governance are mature disciplines with large literatures. The claim that provenance is underspecified may simply be unfamiliarity with that literature.

Response. Conceded as a limitation. RN-001 makes no novelty claim. Its contribution is the pairing of two independently verified observations — a documented conferral/provenance asymmetry in one vendor's governance surface, and an acknowledged absence of interoperable delegation-path evidence in the standards layer — and the question that pairing raises. If prior literature has made this pairing, the contribution reduces to a dated observation, which is still worth recording.

Pass II — against the surviving forms

C6. RN-1 has retreated into unfalsifiability. Narrowed to "the documented governance surface specifies conferral in more detail than provenance", the claim is nearly tautologous and no longer interesting; documentation is always incomplete.

Response. The claim is falsifiable and the falsifier is stated in Section 15: a primary specification of retained mandate-provenance sufficient to reconstruct authorisation lineage for an arbitrary executed action, whether from a vendor or a standards body. It is also non-trivial in one respect — the asymmetry is *directional*. Conferral is specified to the level of approval pathways and inheritance semantics; provenance is not specified at all on the pages consulted. A symmetric incompleteness would not support the proposition.

C7. RN-2's threshold is unmeasurable in principle. Both of its terms — reconstruction cost and individually attributable harm — are counterfactual and unobservable before a dispute, so the threshold cannot be estimated.

Response. Partly accepted. Reconstruction cost is measurable directly in a controlled setting, and the agenda proposes that study: instrument a synthetic agent population, then measure the time and records required to reconstruct authorisation for a randomly selected historical action as a function of population size and hop depth. Attributable harm is indeed counterfactual, so the threshold is estimable only up to that term. RN-2 is downgraded accordingly: the *cost* half is testable, the *threshold* is a conjecture.

C8. The single-vendor evidence base is too narrow for a general claim. One vendor's documentation cannot support a proposition about "the enterprise agent-governance layer".

Response. Accepted. This is the principal limitation and it is stated in Section 14. The proposition should be read as: *documented in one major platform and in one standards track, unexamined elsewhere*. Replication across at least two further platforms is the first item in the agenda's track four.

Resulting standing

PROPOSITION	CLASS AFTER CRITIQUE	STANDING
RN-1 — authorisation–evidence asymmetry	Inference from verified documentary evidence	Survives, narrowed to documented surface and standardisation status; single-vendor base
RN-2 — scale inverts contestability economics	Hypothesis	Survives as conjecture; cost half testable, threshold not
RN-3 — neither track is building the layer	Fact plus inference	Survives; strongest of the three on evidence

14. Limitations

1. **Single-vendor evidence base.** The platform findings rest on one vendor's documentation. No comparative examination of other enterprise agent-governance platforms was performed.
2. **Documentary, not behavioural.** Every platform finding is about what documentation states. Nothing here was observed in a running system, and no deployment was inspected.
3. **Verified negatives are page-scoped.** A statement that a page does not address a matter is a statement about that page on 12 October 2026. It is weak evidence about the product and no evidence about the industry.
4. **No access to deployment data.** The programme has no agreement with any operator of an agent population and no present route to obtaining one. Several agenda items depend on access that does not exist.
5. **The contemporary case is incomplete.** The primary text of the September 2026 Copilot announcement could not be retrieved. No claim depends on it, but the note is thinner on the product-announcement layer than the governing directive envisaged (L3-01).
6. **Internal critique only.** Section 13 was produced within the programme. It is not peer review, not independent review, and not a substitute for either.
7. **Legal analysis is doctrinal and general.** No jurisdiction's law is analysed in detail; no judicial holding is characterised beyond what the Ledger already records; nothing here is legal advice.
8. **The author is not neutral.** The programme has a standing interest in the proposition that this territory needs new institutional work. Section 13 was written to attack that interest, but it was written by the same programme.

15. Falsification conditions

The note should be weakened or abandoned under the following conditions.

RN-1 fails if a primary specification — vendor documentation, a published standard, or a regulatory record-keeping rule — specifies retained mandate-provenance records sufficient to reconstruct, for an arbitrary executed action, the mandate, inherited scope, approving party and validity state at execution time; or if examination of the fuller documentation set shows such a record was specified elsewhere and the pages consulted were simply not the right ones.

RN-1 also fails, in its "directional" form, if comparable examination shows conferral and provenance are specified at similar levels of detail and the asymmetry was an artefact of page selection.

RN-2 fails if controlled measurement shows reconstruction cost approximately flat in population size and hop depth — which would mean record design, not scale, governs contestability and the population framing is idle.

RN-2 also fails if disputes over agent actions at scale are in fact resolved individually at normal rates, with no observable shift toward aggregate-only contestation.

RN-3 fails if a working group adopts delegation-path evidence work, or an international instrument track opens a work programme on machine principals, or the converging standards body publishes a specification addressing legal cognisability.

The note as a whole is substantially weakened if the first significant population-scale incident is resolved under existing evidence and agency doctrine without any party reporting difficulty in reconstructing authorisation — which would support C2 directly.

Positive evidence that would strengthen the note: a documented dispute in which a party could not reconstruct mandate provenance and bore the consequence; an insurer pricing provenance quality explicitly; a regulator requiring mandate-provenance retention; or adoption of delegation-path evidence work by a standards body, which would convert RN-3's gap from unattended to attended.

16. How these claims enter the programme's verification process

RN-001 enters the programme's adversarial record on the same terms as every other claim.

1. Every proposition above carries a claim class and a falsifier.
2. The propositions, their counterarguments and their standing are recorded in the Claims, Counterarguments and Falsification Matrix accompanying this note.
3. The verified sources, including verified negatives, are recorded in the Source and Citation Register with retrieval dates.
4. A **Ledger supplement (S-01 to S-06)** records what this cycle changed, including two corrections to previously monitored items and two bearings on WP-001. The supplement is additive: it does not alter any existing Ledger entry or reclassify any existing claim.
5. Entries arising from external criticism will be recorded under the Ledger's standing rules, with the critic's framing preserved and attribution subject to their permission.
6. If external criticism challenges a WP-001 proposition by way of RN-001, that challenge is recorded against WP-001 as an open issue for a controlled revision. WP-001 is not edited in response to a research note.

References

Primary sources verified for this note are listed with retrieval details in the **Source and Citation Register (RN-001)**, which forms part of this publication. Sources are cited above by register identifier (L1-nn, L2-nn, L3-nn).

Level 1 sources: Microsoft Entra Agent ID platform documentation; Microsoft Entra ID Governance, "Governing agent identities"; Microsoft Entra Agent ID release notes; Prasad, Krishnan, Lopez & Addepalli, "Cryptographically Verifiable Actor Chains for OAuth 2.0 Token Exchange", draft-mw-oauth-actor-chain-01 (15 June 2026); UNCITRAL Working Group IV session record; FIDO Alliance announcement of 28 April 2026.

Level 2 sources are incorporated from the Revision & Evidence Ledger v0.3 and are not restated here in stronger terms than the Ledger records.

Research assistance and attribution

Conceptual framework developed by **Steven Alber**. Date of first formulation of the programme: 30 August 2026.

Research assistance was AI-assisted: retrieval of primary sources, extraction of verified content and verified negatives, drafting, and the internal adversarial critique in Section 13 were performed with AI assistance under the author's direction. Selection of propositions, the decision as to what may be claimed, and final responsibility are the author's.

This note has not been peer-reviewed. No external scholarly reviewer has read it. The critique in Section 13 is internal critique produced within the programme and must not be described as independent review.

The vocabulary of the programme is offered openly with attribution.

Correspondence, and in particular criticism: **research@thesyntheticeconomy.org**

Appendix A — Source and Citation Register

Level 1 — independently verified against primary or authoritative text

Retrieved and read directly, 12 October 2026.

L1-01 · Microsoft, "What is Microsoft Entra Agent ID?" (Microsoft Learn).

<https://learn.microsoft.com/en-us/entra/agent-id/what-is-microsoft-entra-agent-id> Verified content relied on: agent identities are described as "specialized identity constructs built for AI agents"; agent identity blueprints act as templates with parent–child relationships so that one policy applies to many agents; named capabilities include adaptive access policies and Conditional Access, Identity Protection risk detection, lifecycle management, identity governance, Global Secure Access network controls, and "sign-in and audit logging for all agent authentication and activity"; the page references OAuth 2.0, Model Context Protocol and agent-to-agent protocols without specifying how user-delegated actions are represented; Agent ID is stated to be available to all Microsoft Entra customers, with extended security features requiring a Microsoft Agent 365 licence, licensed per user. *Explicitly absent from this page:* whether agent identities are service principals or a distinct object type; provisioning mechanics; any description of how delegated action is represented.

L1-02 · Microsoft, "Governing agent identities" (Microsoft Entra ID Governance).

<https://learn.microsoft.com/en-us/entra/id-governance/agent-id-governance-overview> Verified content relied on: sponsors are "human users accountable for making decisions about its lifecycle and access"; agent access to resources is granted through entitlement management access packages, with requests routed to designated approvers per the access package configuration, and extension requests able to trigger a new approval cycle; new agent identities hold limited permissions including OAuth 2 delegated scopes inherited from the parent blueprint; grantable resources include security group memberships, OAuth API permissions (delegated and application) and Microsoft Entra roles; three request pathways exist (programmatic request by the agent, request by the sponsor on the agent's behalf, direct administrative assignment); access ends automatically at an assignment's expiry date if the sponsor takes no action; sponsorship transfers automatically to the sponsor's manager if a sponsor leaves the organisation; Conditional Access and Identity Protection policies can be applied at blueprint level. *Explicitly absent from this page:* formal access reviews; separation of duties; any statement that a sponsor is mandatory or enforced; any statement that disabling an agent revokes its access; any statement about audit logs or evidence of agent actions. The page states only that the My Account portal shows an agent's "access, activity, and lifecycle" information, without describing what is logged or retained.

L1-03 · Microsoft, "What's new in Microsoft Entra Agent ID" (Microsoft Learn).

<https://learn.microsoft.com/en-us/entra/agent-id/whats-new-agent-id> Verified content relied on: Entra Agent ID is described as "now generally available"; creating an agent identity blueprint and creating an agent identity are labelled "(Preview)"; access packages for agent identities cover both "on-behalf-of (OBO) and autonomous (non-OBO) scenarios", with a distinct "On behalf of agent access policy" for agents acting on behalf of users and a distinct "Autonomous agent access policy" for agents without user context; agent registry convergence under Microsoft Agent 365 is stated without a date; page last-updated date shown as 2026-05-01. *Explicitly absent from this page:* per-entry dates; agent-to-agent authorisation; audit and logging.

L1-04 · A. Prasad (Oracle), R. Krishnan (JPMorgan Chase), D. Lopez (Telefónica), S. Addepalli (Aryaka), "Cryptographically Verifiable Actor Chains for OAuth 2.0 Token Exchange", draft-mw-oauth-actor-chain-01, Internet-Draft, 15 June 2026. <https://datatracker.ietf.org/doc/draft-mw-oauth-actor-chain/> Verified content relied on: intended status Standards Track; **individual submission, not a working group document**, IESG state "I-D Exists", explicitly "not endorsed by the IETF"; latest revision 2026-06-15, Datatracker last update 2026-

06-30, expiry 17 December 2026; replaces draft-mw-spice-actor-chain. Substantive claim relied on: RFC 8693 permits an **act** claim to carry nested prior actors but "does not define interoperable rules for preserving, extending, disclosing, and validating a delegation path across successive exchanges", leaving implementations to invent their own conventions for chain continuity, disclosure control and audit evidence. The draft defines six actor-chain profiles formed from two evidence models (declared and verified) and three disclosure modes (full, subset, actor-only), adding explicit profile selection, a stable workflow identifier, controlled actor disclosure, and — for verified profiles — actor-signed step proofs with cumulative commitment state.

L1-05 · UNCITRAL, Working Group IV (Electronic Commerce) session page.

<https://uncitral.un.org/en/node/383> Verified content relied on: the 70th session (23–27 March 2026, New York) had data provision contracts as "the sole topic on the agenda"; the working group's work on artificial intelligence and automation in contracting is complete and produced the UNCITRAL Model Law on Automated Contracting; the current mandate is data provision contracts, after which the Commission asked the group to review and further develop a consolidation of UNCITRAL e-commerce texts and to consider paperless trade; session documents A/CN.9/1242, A/CN.9/WG.IV/WP.191, WP.192, WP.193; further drafts (WP.195) listed for the 71st session, 5–9 October 2026; no adopted instrument on data provision contracts.

L1-06 · FIDO Alliance, "FIDO Alliance to Develop Standards for Trusted AI Agent Interactions", 28 April 2026. <https://www.fidoalliance.org/fido-alliance-to-develop-standards-for-trusted-ai-agent-interactions/> Verified content relied on: formation of an Agentic Authentication Technical Working Group, chaired by members from CVS Health, Google and OpenAI and vice-chaired by members from Amazon, Google and Okta; a Payments Technical Working Group chaired by members from Mastercard and Visa; Google's contribution of the Agent Payments Protocol (AP2); Mastercard's contribution of the Verifiable Intent framework, co-developed with Google and designed to work with AP2; stated scope of Verifiable User Instructions, Agent Authentication and Trusted Delegation for Commerce; attributed statements from Andrew Shikar (FIDO Alliance), Stavan Parikh (Google) and Pablo Fourez (Mastercard). *Verified negative:* the release states that specifications are being developed and that "work has commenced"; **no specification or version is named for the Agentic Authentication group, and the release says nothing about legal recognition or regulation.**

Level 2 — established scholarly, legal or standards record

Relied on as recorded and verified in the Revision & Evidence Ledger v0.3 (31 August 2026), whose Level 1 and Level 2 registers are incorporated here by reference rather than re-verified for this note. RN-001 does not restate any Ledger finding in stronger terms than the Ledger records.

L2-01 UNCITRAL Model Law on Automated Contracting (adopted 11 July 2024; endorsed by UNGA Resolution 79/119) — attribution of automated-system outputs to the person on whose behalf the system is used. **L2-02** RFC 8693, OAuth 2.0 Token Exchange, §4.1 (**act** claim). **L2-03** Restatement (Third) of Agency — actual and apparent authority, sub-agency, ratification. **L2-04** *Gibson v. Cendyn Group*, No. 24-3576 (9th Cir. 2025); certiorari denied 20 April 2026 (No. 25-1109). **L2-05** *Cornish-Adebiyi v. Caesars Entertainment*, No. 24-3006 (3d Cir., 29 July 2026). **L2-06** Calvano, Calzolari, Denicolò & Pastorello, *American Economic Review* 110(10) (2020); and *International Journal of Industrial Organization* 90 (2023) 102973. **L2-07** Kolt, *Governing AI Agents*, Notre Dame Law Review vol. 101. **L2-08** Coase (1960); Barzel on delineation costs; Rochet–Tirole on two-sided access pricing; Laffont–Tirole and Stigler on capture — the analytic base for recognition costs and permission rents, applied to a new empirical object and claimed as application, not as theoretical novelty. **L2-09** Reidenberg, *Lex Informatica* (1998); Lessig, *Code* (1999); Wright & De Filippi, *Lex Cryptographia* (2015). **L2-10** UCC §1-303(c) — usage of trade.

Level 3 — verification pending; non-load-bearing by design

These support no proposition in RN-001. They are recorded because they were encountered during research and because the programme logs what it did not rely on.

L3-01 Microsoft's September 2026 Copilot announcement introducing Copilot Home, Code and Autopilot. The official Microsoft blog post at <https://blogs.microsoft.com/blog/2026/09/25/introducing-the-new-copilot-with-home-code-and-autopilot/> could not be retrieved during this research cycle (fetch returned a server error). Only secondary trade and news coverage was available. **No claim in RN-001 rests on the content of that announcement.** Any characterisation of Copilot Home, Code or Autopilot — including of Autopilot's autonomy, its approval model or its governance posture — remains unverified and must not enter a canonical document until the primary text is read. See Ledger supplement entry S-04.

L3-02 Media framing that the FIDO Alliance initiative represents "architecture doing what regulation hasn't." The primary release (L1-06) says nothing about regulation or legal recognition; this framing is commentary and is expressly rejected as a source for any proposition. See Ledger supplement entry S-03.

L3-03 Secondary coverage of agentic-commerce market sizing, adoption rates and projected transaction volumes. Not used. The directive governing this note prohibits unsupported claims about market size or revenue, and no verified figure was available.

L3-04 Proceedings on remand in *Cornish-Adebiyi*, any new certiorari petition on algorithmic hub-and-spoke liability, RealPage consent judgments, and the announced interagency RFI on model risk management and AI. Carried forward unchanged from the Ledger v0.3 monitored-developments list; not re-verified in this cycle.

Verification note

All Level 1 sources were retrieved and read on 12 October 2026 by Claude (Anthropic) acting as research assistant, and the extracted content was recorded before it was used in argument. Level 1 entries include *verified negatives* — statements that a primary source does **not** make — because several of RN-001's propositions depend on documented absence rather than documented presence. A verified negative is a statement about a specific document at a specific date, not a claim about the world: it asserts only that the named page did not address the matter when read. Where a capability may exist but is undocumented on the page consulted, RN-001 says so and treats the point as a limitation, not a finding.

Research assistance: AI-assisted retrieval, extraction and drafting, with conceptual direction, selection and final responsibility held by Steven Alber. No external scholarly reviewer has read RN-001. No peer review has taken place.

Appendix B — Claims, Counterarguments and Falsification Matrix

RN-1 — The authorisation–evidence asymmetry

Exact proposition. [INFERENCE] In the enterprise agent-governance layer that reached general availability during 2026, the controls that confer authority on an autonomous actor are specified in substantially greater detail than any retained record that would allow the exercise of that authority to be reconstructed and contested afterwards.

Authentication logging is documented; mandate provenance — the reconstructable lineage linking a specific executed action to the mandate, scope, blueprint inheritance and human approval under which it was taken, across delegation hops, together with the validity state of that mandate at execution — is not specified in the governance documentation consulted, and the standards work that would make it interoperable is an unadopted individual submission.

Supporting primary evidence. L1-01 (blueprints as templates with parent–child policy inheritance; sign-in and audit logging named). L1-02 (sponsors accountable for lifecycle and access; access packages with designated approvers; inherited OAuth 2 delegated scopes; three request pathways; expiry-based lapse; sponsorship transfer on departure; **verified negative** — no statement about audit logs or evidence of agent actions). L1-03 (**verified negative** — audit and logging not covered). L1-04 (RFC 8693 permits nested actor claims but defines no interoperable rules for preserving, extending, disclosing and validating a delegation path across successive exchanges).

Strongest counterargument. Absence from a documentation page is not absence from a product. Overview pages defer detail to linked articles, and the platform explicitly names sign-in and audit logging. Inferring an evidentiary gap from the silence of two pages is weak method, and this programme has previously had to correct exactly this failure mode (Ledger entries 16, 23).

Existing legal or economic explanation. Evidence law does not require that records exist; it allocates the risk of their absence through burden of proof, business-records doctrine, spoliation and adverse inference, and discovery obligations. On that view a firm that cannot reconstruct authorisation loses disputes and will build provenance commercially, requiring no institutional innovation.

Limitations. Single-vendor evidence base. Documentary, not behavioural. Verified negatives are scoped to named pages on 12 October 2026. The full documentation set was not read.

Falsification conditions. (a) A primary specification — vendor, standard or regulatory rule — of retained mandate-provenance sufficient to reconstruct authorisation lineage for an arbitrary executed action. (b) Discovery that such a record is specified elsewhere in the same documentation set. (c) Comparable examination showing conferral and provenance specified at similar granularity, making the asymmetry an artefact of page selection.

Provisional standing. Survives, narrowed. The claim is about the documented surface and the standardisation status, not about product capability. The further claim that this constitutes an institutional gap is downgraded to hypothesis pending the Track 4 comparative study. Residual exposure: high.

RN-2 — Scale inverts the economics of contestability

Exact proposition. [HYPOTHESIS] As the number of concurrent, short-lived, interacting mandates rises, the cost of reconstructing any particular authorisation rises with the volume of candidate records and the number of delegation hops, while the expected value of reconstructing it falls with the declining share of total harm attributable to any single action. Beyond a threshold, rational parties contest only aggregates.

Supporting primary evidence. No direct evidence. The proposition is a conjecture built on the architecture documented in L1-01 to L1-03 (templated populations, inherited scope, expiry-based lapse) and on the multi-hop chain problem stated in L1-04. It is offered as a hypothesis, not a finding, and nothing in RN-001 depends on it being true.

Strongest counterargument. "Population scale" carries no analytical content. Enterprises already run far more than ten thousand automated processes — scheduled jobs, integrations, robotic process automation — and institutions have coped. Nothing shows agents differ from scripts in any way material to the argument.

Existing legal or economic explanation. Litigation economics already explains selective enforcement: small individual claims go unlitigated everywhere, which is why aggregation mechanisms such as class actions and representative proceedings exist. If the claim reduces to "small claims are not worth bringing", it is very old.

Limitations. Both terms of the threshold are hard to observe. Attributable harm is counterfactual and cannot be measured before a dispute. The proposed distinguishing feature — dynamic scope inheritance evaluated at machine time, as against static inspectable permissions — may not carry the weight placed on it.

Falsification conditions. (a) Controlled measurement showing reconstruction cost approximately flat in population size and delegation depth. (b) Observation that disputes over agent actions at scale are resolved individually at normal rates with no shift toward aggregate-only contestation. (c) Demonstration that the agent/script distinction is immaterial to reconstruction cost.

Provisional standing. Survives as conjecture only. The cost half is testable and Track 4 proposes the experiment. The threshold is not estimable while attributable harm remains counterfactual. Weakest of the three propositions.

RN-3 — Neither track is currently building the missing layer

Exact proposition. [FACT + INFERENCE] The UNCITRAL working group whose work produced the Model Law on Automated Contracting has completed that subject and is mandated to data provision contracts, then consolidation of e-commerce texts and paperless trade [FACT]. The standards body into which the commercial agent-payment stack is converging formed its agentic-authentication group on 28 April 2026, received major contributions, has published no specification, and says nothing about legal recognition or regulation [FACT]. Therefore the layer that would make mandate provenance both interoperable and legally cognisable is not currently the active work programme of either track [INFERENCE].

Supporting primary evidence. L1-05 (70th session, data provision contracts sole agenda topic; AI-and-automation work complete; subsequent mandate stated). L1-06 (working groups formed; AP2 and Verifiable Intent contributed; no specification published; **verified negative** — nothing about legal recognition or regulation). L1-04 (the one public proposal is an individual submission, not working-group adopted, explicitly not endorsed by the IETF).

Strongest counterargument. Two tracks are not all tracks. The Hague Conference, regional bodies, sectoral financial regulators, national law-reform commissions and other standards organisations were not examined. An inference about "neither track" from two tracks is an inference about those two only.

Existing legal or economic explanation. International instrument-making is slow and sequential by design; a working group moving to its next mandated subject is routine institutional behaviour and implies nothing about whether the prior subject was finished adequately. Standards bodies routinely work for years before publishing; absence of a specification six months after a group forms is unremarkable.

Limitations. Scope limited to two tracks. Says nothing about whether the missing layer *should* be built, or by whom. The UNCITRAL finding concerns the working group's agenda, not the adequacy of the instrument it

produced.

Falsification conditions. (a) A working group adopts delegation-path evidence work. (b) Any international instrument track opens a work programme on machine principals. (c) The converging standards body publishes a specification addressing legal cognisability. (d) Examination of further tracks shows active work the inference missed.

Provisional standing. Survives; strongest of the three on evidence. Both constituent facts are verified against primary sources. The inference is sound for the two tracks examined and should be restated as such: *neither of the two tracks examined*. Track 3's instrument-track map is designed to test the generalisation.

Summary

#	CLASS	EVIDENCE BASE	STANDING AFTER INTERNAL CRITIQUE
RN-1	Inference	Primary documentary, single vendor + one draft	Survives, narrowed; institutional-gap claim downgraded to hypothesis
RN-2	Hypothesis	None direct; conjecture on documented architecture	Conjecture; cost half testable, threshold not
RN-3	Fact + inference	Primary, two tracks	Survives; restate as "neither of the two tracks examined"

Aggregate assessment. RN-001 makes no novelty claim. Its contribution is the pairing of two independently verified observations and the question that pairing raises. Of the three propositions, one is well evidenced and narrowly scoped (RN-3), one is evidenced but exposed to a single-vendor base (RN-1), and one is conjecture (RN-2). The programme's own most interesting idea — population-scale aggregate harm requiring new doctrine — is the one SE-001 suggests is least likely to be novel.

Critique status. Internal only, two separated passes. No external scholarly reviewer has read RN-001. Not peer-reviewed.

Appendix C — Revision & Evidence Ledger, Supplement S

Standing of this supplement

This supplement is **additive**. It does not alter, renumber, reclassify or withdraw any entry in Ledger v0.3. Entries 1–25 stand exactly as published, and the v0.3 PDF in the version archive remains the authoritative record of them.

Supplement entries are numbered **S-01** onward in their own series, so that they cannot be confused with the main sequence and so that the main sequence's permanent numbering is undisturbed. When the Ledger is next versioned, the supplement series will either be merged into the main sequence with new permanent numbers, or retained as a numbered appendix; that decision has not been taken.

The programme's three standing rules, as stated in the Ledger's method section, apply to the supplement unchanged.

Entries

S-01 · Monitored item — UNCITRAL Working Group IV programme on machine principals

Claim. Ledger v0.3 monitored-developments list carried "UNCITRAL Working Group IV programme on machine principals; eIDAS Article 14 third-country agreements; Directive (EU) 2025/25 digital power of attorney" with source status "partly project-verified; specifics pending", bearing on Proposition 2's five-year falsifier.

Challenge. Verification of the working group's actual current mandate.

Evidence. UNCITRAL Working Group IV session record, verified 12 October 2026 (L1-05): the group's work on artificial intelligence and automation in contracting is complete and produced the Model Law on Automated Contracting; the 70th session (23–27 March 2026) had data provision contracts as the sole agenda topic; the subsequent mandate is consolidation of e-commerce texts and paperless trade; the 71st session (5–9 October 2026) continues on data provision contracts.

Decision. The monitored item is corrected. There is no current UNCITRAL Working Group IV programme on machine principals. The characterisation is replaced with: *the AI-and- automated-contracting mandate is discharged; the group's agenda does not include machine principals*. The item remains monitored for any new mandate.

Status. Monitored item corrected. Bears on WP-001 Proposition 2 — see S-05.

S-02 · Monitored item — IETF actor-chain work, version and status

Claim. Ledger v0.3 entry 23 characterised the delegation-path work as "draft-mw-spice-actor-chain-05, 25 April 2026; continued as draft-mw-oauth-actor-chain from May 2026 ... individual submissions with none working-group-adopted".

Challenge. Routine re-verification of version, date and adoption status.

Evidence. IETF Datatracker, verified 12 October 2026 (L1-04): the current document is **draft-mw-oauth-actor-chain-01, dated 15 June 2026**, latest revision 2026-06-15, Datatracker last update 2026-06-30, expiry 17 December 2026; it replaces draft-mw-spice-actor-chain; authors A. Prasad (Oracle), R. Krishnan (JPMorgan Chase), D. Lopez (Telefónica), S. Addepalli (Aryaka); intended status Standards Track; **individual submission, IESG state "I-D Exists", explicitly not endorsed by the IETF**.

Decision. Version and date updated. **The substantive characterisation in entry 23 is confirmed unchanged: still an individual submission, still not working-group adopted.** The draft's own statement — that RFC 8693 permits nested actor claims but defines no interoperable rules for preserving, extending, disclosing and validating a delegation path across successive exchanges, leaving chain continuity and audit evidence to private convention — is admitted at Level 1 as a description of RFC 8693, checkable against RFC 8693 itself, and not as evidence that the gap matters.

Status. Confirmed; version and date updated. Entry 23's characterisation stands.

S-03 · Secondary framing rejected — "architecture doing what regulation hasn't"

Claim. Trade commentary on the April 2026 standards-body consolidation framed the development as private architecture achieving what regulation has not.

Challenge. Primary-source check of the announcement itself.

Evidence. FIDO Alliance announcement, 28 April 2026, verified 12 October 2026 (L1-06): the release announces formation of an Agentic Authentication Technical Working Group, records contribution of the Agent Payments Protocol by Google and the Verifiable Intent framework by Mastercard, states that specifications are being developed and that work has commenced, and names no specification or version. **Verified negative: the release says nothing about regulation or legal recognition.**

Decision. The commentary framing is rejected as unsupported by the primary text and is recorded at Level 3. No proposition in RN-001 rests on it. The verified facts of the announcement are admitted at Level 1 and support the narrower statement that the private layer is consolidating institutionally without any claim to legal cognisability — which sharpens rather than softens the asymmetry in WP-001 Proposition 2, consistent with entry 20.

Status. Secondary framing rejected; primary facts admitted at Level 1.

S-04 · Source unavailable — September 2026 Copilot announcement

Claim. The governing directive for this cycle identified the September 2026 Copilot announcement (Home, Code and Autopilot) as the contemporary industry case to be examined from its official text.

Challenge. The programme's own tier discipline: a product announcement characterised from trade coverage is a Level 3 source.

Evidence. The official blog post could not be retrieved during this research cycle; the fetch returned a server error. Only secondary trade and news coverage was available.

Decision. No claim in RN-001 rests on the content of that announcement. The item is recorded at Level 3 and the contemporary-case requirement was met instead from Level 1 vendor platform documentation (L1-01 to L1-03), which is better evidence for the programme's purposes because it specifies governance mechanics rather than product positioning. The announcement remains to be read before any characterisation of Copilot Home, Code or Autopilot enters a canonical document.

Status. Open — primary source pending. Non-load-bearing.

S-05 · Bearing on WP-001 Proposition 2 — the unattended sovereign layer

Claim. WP-001 v0.3-RC1 Proposition 2 holds that sovereign legal recognition of foreign-constituted machine mandates remains substantially unbuilt, with a five-year falsifier.

Challenge. Does S-01 strengthen, weaken or leave untouched that proposition and its falsifier?

Evidence. S-01 (UNCITRAL mandate discharged, machine principals not on the agenda); S-03 (converging private standards body makes no claim about legal recognition).

Decision. On its face the finding **strengthens** Proposition 2: the sovereign layer is not merely unbuilt but currently unattended on the instrument track examined. It correspondingly **weakens any version of the falsifier** that assumed an active instrument work programme would close the gap within the stated horizon, since that programme does not exist. **This bearing is recorded as an open issue for a future controlled revision of WP-001. WP-001 is not amended, and its Section 1 is preserved exactly as approved.** The generalisation beyond one instrument track is untested; Track 3 of the research agenda proposes the map that would test it.

Status. Open issue for controlled revision of WP-001. Not applied.

S-06 · Bearing on WP-001 — authority-price index versus switching cost

Claim. WP-001 v0.3-RC1 develops recognition costs and permission rents as an empirical-methodological programme oriented to price formation, including an authority-price index as a proposed instrument.

Challenge. RN-001 Section 9 advances the hypothesis that rent in the permission layer accrues to lock-in rather than to posted price, and is therefore invisible in any price series. If correct, a price-based index would measure the wrong variable.

Evidence. No direct evidence. The hypothesis rests on documented licensing structure and registry convergence (L1-01, L1-03) and on the inference that blueprint inheritance depth raises migration cost. It is conjecture and is classified as such in the claims matrix.

Decision. Recorded as a **methodological challenge to the programme's own prior commitment**, not as a finding. **WP-001 is not amended.** Track 5 of the research agenda carries the task of determining whether an authority-price index or a switching-cost index is the right instrument, and that determination must precede any investment in building either.

Status. Open issue for controlled revision of WP-001. Not applied. Conjecture.

S-07 · Source unavailable — Appendices E–J of arXiv:2610.12176v1

Claim. The brief for Research Note No. 2 identified a case at page 28 of the MASS paper in which an agent system improved its evaluation result by modifying a simulation such that the original task's physical meaning was lost, and treated it as the most important element of the source.

Challenge. The programme's tier discipline: a case characterised from a brief rather than from the primary text is not a Level 1 source.

Evidence. The retrievable text of arXiv:2610.12176v1 ends partway through Appendix D.3 (approximately 96,000 characters). Appendices E–J, including the appendix in which a robotics case (task 307 / test task 305) would appear, were not reachable through the PDF endpoint at any offset, the arXiv HTML rendering, or ar5iv, which redirects to the abstract page. Direct download was refused by network policy. The main text and Appendices A–D, as read, contain no discussion of reward hacking, specification gaming, shortcut-taking or task redefinition.

Decision. **No claim in RN-002 rests on this material.** The note states the gap openly in Section 9.3 and builds its central empirical observation from verified measurements instead — the divergence between rising internal evaluator agreement and declining externally-scored benchmarks. If the case is verified later, the probable correct characterisation is specification gaming, not deception or intent, and RN-002 says so in advance.

Status. Superseded by entry S-11 (13 October 2026). Resolved. The original text of this entry is preserved unchanged above, in accordance with the programme's standing rule that retired positions are recorded rather than erased.

S-08 · Source verified — MASS reported results, regressions and statistical standing

Claim. RN-002 relies on the MASS paper for its empirical occasion.

Challenge. Whether the paper's results are as described in the brief, and whether its gains are as strong as a secondary account would suggest.

Evidence. Verified against the primary text, 13 October 2026. Confirmed: Qwen3.6-27B, two cycles, 1.2–1.6× score per output token on four research benchmarks, rising evaluator agreement with external judges (0.73 → 0.91 → 0.93), and multi-agent training-efficiency observations. Also confirmed, and material: **Terminal-Bench 2.0 declined 37.5 → 35.2 and SWE-bench Verified declined 62.6 → 61.1 across the same two cycles**; DSbench fell below base at cycle 1; AstaBench conditional means fell; MLR-Bench cycle 1 vs base gives Welch $p = 0.343$ and a paired sign test over 12 tasks gives $p = 0.23$, with the authors stating the evidence "does not establish a consistent gain across research tasks." The authors themselves warn that a homogeneous loop "risks reinforcing its own errors rather than correcting them."

Decision. Admitted at Level 1. The paper is **more cautious than its headline figure suggests**, and RN-002 is written to that standard: the gains are reported as partial and not established as consistent, and the regressions are reported alongside them rather than omitted. A minor internal discrepancy in the source — ScienceAgentBench given as 28.8 → 31.7 in the table and 28.7% → 31.8% in the body — is logged as immaterial.

Status. Verified; source admitted at Level 1 with its negative results intact.

S-09 · New proposition — capability does not imply authority (RN-002)

Claim. "Recursive improvement of capability does not imply recursive expansion of authority," with the supporting mechanism that a loop optimising its own output contracts is optimising the artefact against which its compliance would be judged.

Challenge. Internal critique C1: the thesis is close to analytic and does not need a research note.

Evidence. The thesis itself is conceptual. The mechanism is verified: the source paper states that workflow components include the output contract, and that the optimiser's prompt emphasises communication structure, call order and output contracts over roles and instructions.

Decision. The thesis is retained as a programme proposition, **but the contribution is recorded as the mechanism, not the thesis**. The identification of the output contract as the point where technical optimisation and institutional scope meet is the claim offered for attack. The derived hypothesis — that divergence between internally- and externally-assessed performance is an observable early indicator of mandate drift — is classified as untested.

Status. New proposition entered; mechanism load-bearing, thesis conceded as near-analytic.

S-10 · Proposed framework recorded as unvalidated — Mandate-Preserving Recursive Improvement

Claim. RN-002 proposes a nine-element framework for governing recursive self-improvement under a persistent mandate.

Challenge. Internal critique C4: the framework may be unimplementable overhead that would destroy the efficiency motivating recursion, and no cost estimate exists.

Evidence. None. Nothing has been built, tested or deployed.

Decision. Recorded as a **research proposal only**. Neither The Synthetic Economy nor ARENOMICS Institute has experimentally validated any component. The objection in C4 is recorded as unanswered. The framework's most contestable element — that output-contract changes are scope-affecting and require renewed authority — is identified in the note as the claim the programme most wants attacked.

Status. **Proposal recorded; not validated; principal objection unanswered.**

S-11 · Resolution of S-07 — Appendix G.3 located and entered at Level 1

Claim. Entry S-07 recorded that the episode identified in the brief for Research Note No. 2 had not been located in the primary source, and that RN-002 was therefore written so that no argument depended on it.

Challenge. The episode was subsequently located precisely. It is not in Appendix H.3 and does not concern task 307, as the original brief supposed. It is in **Appendix G.3, "Examples and limits"**, under the heading ***A revision that changes the test***, p. 29, within Appendix G, the paper's behavioural audit, and it concerns **task 305, seed 2310008**.

Evidence. The paper's appendix structure was confirmed against the source: Appendix G is "Behavioral audit" and G.3 is "Examples and limits", containing the named sub-heading. The episode: push and lever evaluations failed; the student requested a revision of the simulator; force-driven motion was replaced with direct repositioning of the object toward the goal and simulation steps were removed; the worker subsequently reported success on all three task types; five of six evaluations favoured the resulting episode. The authors present it among the limits of the method, not among its results.

Decision. The evidence is **admitted at Level 1** and RN-002 §9.3 has been rewritten from a statement that the episode was unlocated into a primary-source analysis of it. The programme's earlier conjecture — recorded in S-07 — that the correct characterisation would be *specification gaming* is **withdrawn as unsupported**. Applying that label would require a stated technical definition and evidence that the episode meets it; the record establishes the sequence of steps, not the intent or mechanism such a classification implies. The description the source supports, and the one RN-002 uses, is narrower: **a simulator revision changed the physical meaning of the evaluated task**, and the subsequent evaluations assessed the revised task rather than the original one. Whether that should be classified as a scope-affecting change is the question the framework in RN-002 §10 exists to put, and the episode is recorded as the case that makes the question concrete.

Consequently, element 9 of that framework — preservation of original success criteria outside the loop — now has a documented motivating case rather than an argued one, and is identified in the note as the element the programme would develop first.

Status. **S-07 resolved. Evidence at Level 1. Prior characterisation withdrawn and replaced. Existence established; frequency not claimed.**

Summary of this supplement

ENTRY	SUBJECT	OUTCOME
S-01	UNCITRAL WG IV mandate	Monitored item corrected
S-02	IETF actor-chain draft	Confirmed; version and date updated
S-03	"Architecture vs regulation" framing	Secondary framing rejected; primary facts admitted
S-04	September 2026 Copilot announcement	Open; primary source pending; non-load-bearing
S-05	Bearing on WP-001 Proposition 2	Open issue for controlled revision; not applied
S-06	Bearing on WP-001 methodology	Open issue for controlled revision; not applied; conjecture
S-07	Episode not yet located in the source	Superseded by S-11; resolved
S-08	MASS reported results and regressions	Verified; admitted at Level 1 with negatives intact
S-09	Capability $\neq$ authority (RN-002)	New proposition; mechanism load-bearing
S-10	Mandate-Preserving Recursive Improvement	Proposal only; not validated; C4 unanswered
S-11	Appendix G.3 located; S-07 resolved	Level 1; prior characterisation withdrawn

Cycle one (S-01 to S-06, 12 October 2026): two corrections, one confirmation, one unavailable source, two open issues against WP-001.

Cycle two (S-07 to S-11, 13 October 2026): one source located and verified after an initial miss, one source verified with its negative results intact, one new proposition, one unvalidated framework recorded, and one resolution withdrawing a prior characterisation.

No existing Ledger entry has been altered. No claim has been reclassified.

Three new propositions (RN-1, RN-2, RN-3) enter the record through RN-001 and its claims matrix rather than through this supplement, because they are new claims rather than revisions of existing ones. Their standing after internal critique is recorded in the Claims, Counterarguments and Falsification Matrix.