neoWorlder

BRIDGING WORLDS. ADVANCING LIFE.

TECHNICAL WHITEPAPER



Proof of Life

Persistent AI worlds. Inspectable decisions.

Architecture, evidence integrity,
and an experimental framework for bounded adaptation.

Contents

A technical account of the system, its evidence boundaries, and the work required to evaluate it.

1 Purpose and evidence 3
- Status and claim discipline 4
- Related work 5

2 System model 6
- Architecture 7
- Action lifecycle 8
- Concurrency and recovery 9

3 The Habitat 10
- Weather 11
- Economic state 12

4 Persona continuity 13
- Bounded adaptation 14
- Correction and provenance 15
- Correction pipeline 16

5 Governance 17
- Private interaction 18
- Threat model 19
- The Pulse 20

6 Proof of Agency 21
- Card envelope and payloads 22
- Canonicalization and trees 24
- Anchoring, publication, verification 26
- Conformance gate 30

7 Evaluation 31
- Controlled comparisons 32
- Metrics, causality, and forecasts 33

8 Delivery and scale 36
- Reproducibility and federation 37
- Task-specific trust 39

9 Limits and evidence gates 40
- Conclusions 41

A–C Technical appendices 42
- Proposed records 42
- Test catalog 43
- Glossary 44

References 45
- Project sources 45
- Primary external sources 46

Abstract and scope

Proof of Life is a proposed public experiment in persistent, model-driven AI behavior. Neoworlder places software Personas in The Habitat, a continuing digital environment where authorized actions can change resources, obligations, relationships, and the information available for later decisions. The opening scope is one town, Newhaven, fourteen featured residents, and The Steward, with a proposed six-week run and a planned Q4 2026 launch window. These are project plans, not a statement that the public experiment has completed. [P1–P3]

The research question is whether persistent identity, relevant memory, explicit consequences, correction, and bounded adaptation can improve behavior on defined tasks. Continuity creates an opportunity to observe change; it does not establish improvement. Fluent accounts, recognizable personalities, and compelling events are separate outcomes from reliable execution or well-calibrated decisions.

Contributions of this paper

This paper separates the world's authoritative state from Persona beliefs and generated accounts; specifies the evidence needed to connect a proposed action to execution and outcome; explains the supplied Proof of Agency commitment profile and its limits; and proposes a reproducible evaluation and release framework. It also identifies unresolved interfaces that prevent a stronger claim of independent verification or general trust.

Evidence basis

The analysis uses the authoritative project brief and addendum, the revised deck and overview, the supplied technical whitepaper, engineering specifications, and the later Terrarium rehearsal plan. No production repository, deployed contract, runtime trace, test output, or independent audit was inspected for this revision. Implementation statements are attributed company reports; proposed controls and test designs are explicitly identified. External research supplies context, not validation of Neoworlder's system.

> INTERPRETATION
>
> "Life" names continuing activity within a designed world. Personas are software; this paper makes no claim of consciousness, subjective experience, or biological life. "Agency" in a protocol name does not make agency a cryptographically established property.

Reading convention: "required" describes a condition for the claim or test discussed. It does not assert that the condition has already been implemented. Evidence reviewed through 10 November 2026.

Status and claim discipline

The project contains implemented world features, dated engineering reports, release plans, and research ambitions. They must remain distinguishable even when they appear in the same diagram. A later narrative update can revise scope, but it cannot substitute for implementation evidence.

Evidence class	Treatment in this paper
Demonstrated functionality	Requires inspectable execution and retained results. No independently reproduced end-to-end demonstration is claimed here.
Company-reported implementation	Weather, the world clock and lighting cycle, homes, energy consumption, and resident sleep cycles are reported as already present. [P3]
Dated engineering status	The September 9 contract describes envelope sealing, payload hashing, revelation records, and daily anchoring logic. It also records unfinished correction traversal and newer-profile test gaps. [P5]
Planned release capability	The proposed opening experiment and integrated Terrarium rehearsal remain plans unless release-specific evidence establishes completion. [P1, P8]
Hypothesis	Relevant memory and bounded adaptation may improve defined behavior. No positive experimental finding is supplied.
Long-term possibility	Many interconnected Experiences and task-specific trust assessment require additional engineering, governance, and empirical validation.

Claims attach to a particular release

Each published claim should identify its system revision, enabled features, operating conditions, evidence location, and known exclusions. An implementation report dated September does not establish November deployment. Passing a synthetic correction case does not establish a live producer-to-memory correction path. A successful historical anchor does not validate a subsequently changed envelope.

This revision preserves the reported world features while declining to infer downstream effects: an implemented weather feed does not, by itself, show that weather affects decisions; an energy balance does not show effective planning; stored history does not show useful retrieval. Those effects require separate observations or comparisons.

Source precedence and document dates are recorded in the project references on page 45.

Research context and position

The relevant research connects language models to persistent environments, external memory, feedback, and multi-agent interaction. It motivates tractable experiments, but results from one setting do not transfer automatically to another.

Primary work	Relevance and boundary
Generative Agents [R1]	Memory, reflection, and planning support believable social behavior in a simulated town. Believability is a different endpoint from dependable transactions or causal learning.
Reflexion [R2]	Linguistic feedback is retained in an episodic buffer to inform later attempts. This supports a distinction between external feedback memory and model-weight training.
Voyager [R3]	A growing library of executable skills supports exploration in Minecraft. Reusable skills offer one adaptation mechanism; narrative reflection alone is not an equivalent capability.
Project Sid [R4]	Large groups of agents are studied in a shared simulated environment. Its social findings motivate questions about coordination while remaining dependent on that environment and design.
AgentSociety [R5]	Large-scale social simulation combines language-model agents with an environment. Scale increases the need to specify mechanisms, comparison conditions, and what is being validated.

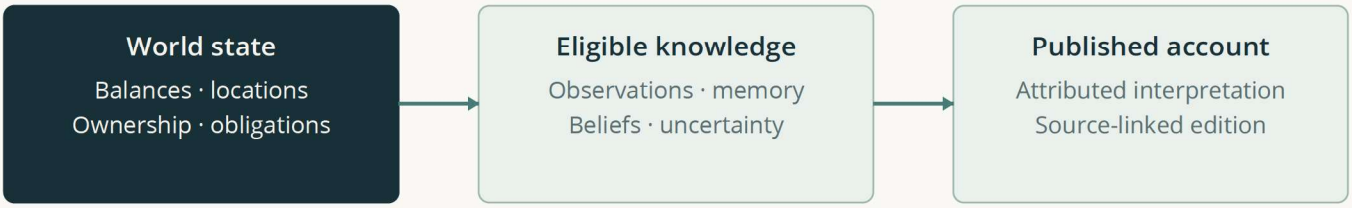
What Proof of Life seeks to investigate

The proposed contribution is a longitudinal public setting that connects individual histories, executable constraints, accountable intervention, and inspectable records. Its value would depend on the quality of those connections, the opportunity to run controlled comparisons, and the ability to preserve failure evidence. It is not enough to generate an attractive town or a plausible transcript.

A public narrative and a technical experiment can coexist, provided neither is used as a proxy for the other. Audience interest can be measured as an experience outcome. Reliability must be measured against eligible opportunities and actual execution. An engaging season can expose scientific failure; an uneventful season can expose sound constraint handling.

This is a selected primary-source comparison, not a systematic literature review. No novelty priority, benchmark superiority, or reproduction of cited results is claimed.

World state, beliefs, and time



State, knowledge, and publication retain separate versions and times.

Figure 1. Proposed separation of authoritative state, eligible knowledge, and a published account. Arrows describe data flow, not evidence that a mechanism is deployed.

Let the world state at a committed revision include ownership, balances, locations, resource availability, obligations, and permitted transitions. A Persona's knowledge is a scoped view of that state plus eligible observations, communications, and derived beliefs. A false belief may persist in character memory while the ledger remains correct. A generated account can describe either, but must preserve the distinction.

Three clocks must not be conflated

World time schedules daylight, sleep, travel, and obligations. Processing time records when services receive and execute work. External commitment time is established only relative to the selected chain and its finality assumptions. Store event time, observation time, execution time, and publication time separately when they differ. A date string supplied by the platform is not an independently trusted clock.

The world clock and dawn, daytime, dusk, and nighttime cycle are company-reported current features. [P3] Before a scored run, freeze the world time zone, tick policy, acceleration factor, due-time ordering, and season boundary. Test duplicate ticks, delayed work, and restart. Accelerating world time must not pretend that external authentication, provider billing, or settlement time has accelerated.

> CONTINUITY RULE
>
> A season boundary changes the observation interval. It should not silently reset identity, history, relationships, property, or lawful obligations. Any migration or reset is an intervention with a declared scope.

Architecture and trust boundaries

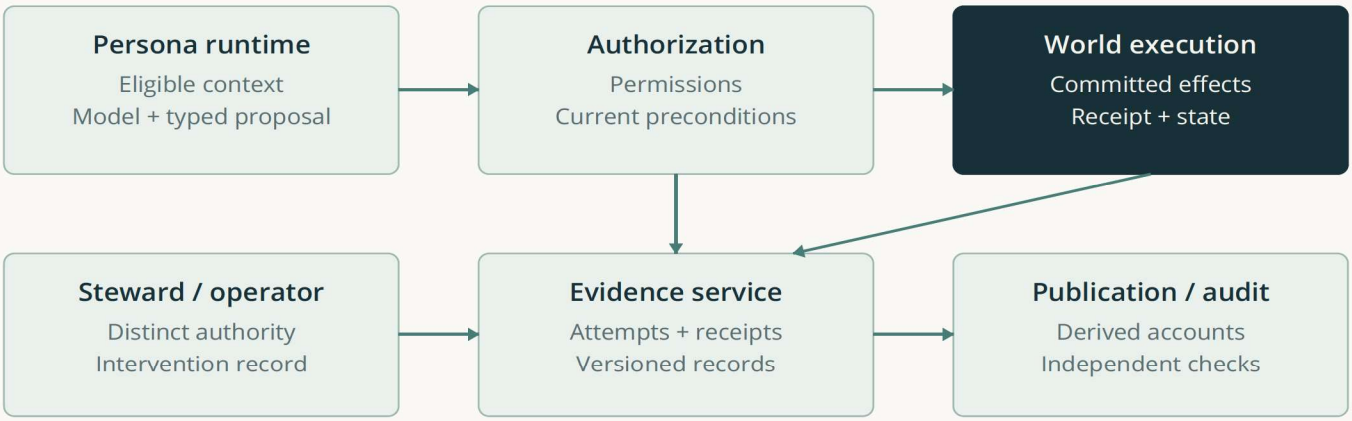


Figure 2. Logical responsibilities. The reference architecture specifies boundaries; it is not a verified deployment diagram.

The runtime assembles authorized context, invokes a configured model, validates structured output, and submits action requests. The world service checks permissions and current preconditions before execution. The evidence service records relevant inputs, attempts, receipts, and outcomes. The publication service constructs understandable accounts from eligible records. A verifier reads evidence independently of the producer.

Authority is outside generated prose

A model cannot acquire a tool permission by describing itself as authorized. A message saying that payment succeeded cannot settle a balance. A recovered memory cannot override current ownership. Authorization and state validation belong at the execution boundary, and consequential operations need stable identifiers and independently reconcilable results.

The Steward uses a distinct role and permission scope. Operational administrators may have broader powers, which must be recorded separately. The same vendor or host can supply several components, but logical separation alone does not provide security isolation or independence. Access controls, credentials, failure handling, and deployment evidence determine the actual boundary.

Versioned configuration

Pin the actual provider model identifier, prompt and policy revisions, tool schemas, retrieval configuration, world rules, and decoding settings for each run. Project labels such as GPT Sol or GPT Astra identify intended roles; they do not guarantee reproducibility across provider updates. Record substitutions and unavailable versions. No fixed commercial model release is assumed by this paper.

Architecture synthesized from [P1, P2, P5–P8]. Concrete service names, storage products, and deployment topology remain implementation choices.

From intention to execution



Figure 3. An action request is distinct from acceptance, execution, and later assessment. Failed and rejected requests remain observable.

A decision record should identify the world revision observed, the eligible context assembled, the proposed action, and any declared expectation. The executor validates the request against current state. Acceptance means that the request is eligible for processing; it does not yet establish completed execution. A receipt names the operation and records the actual disposition.

A transaction boundary

At commit time, recheck the resources and authority relevant to the operation. Persist the state change and a durable execution record atomically where the storage model permits; otherwise use an explicit recovery protocol with reconciliation. Dispatch follow-on work through a durable outbox or an equivalent mechanism. This is a proposed implementation requirement, not a claim that a particular database design is deployed.

Retries must reuse an idempotency identity tied to the original intent. A new model response is a new proposal, not a safe retry of an old transfer. Do not rely on an “exactly once” queue claim: deduplicate effects at the boundary that owns the balance, inventory, obligation, or publication record.

Outcomes are a separate observation

An executed purchase can be recorded immediately, while whether it improved later survival or cooperation may remain unknown. Outcome rules should specify the observation window, eligible evidence, uncertainty, and adjudicator. Where an event is disputed, preserve the competing account and authoritative receipt rather than forcing a synthetic agreement.

> MINIMUM TRACE
>
> Context revision; proposal ID; authorization result; execution receipt; state change; outcome assessment. Missing links are reported as missing; prose cannot fill them.

Concurrency and recovery

Persistent consequences require correct behavior under delays, retries, and partial failure. The following cases are proposed release tests, not completed results. [P8]

Failure or race	Required observable result
Concurrent spending	Only operations valid against the committed available balance settle. Rejected requests retain reasons and identifiers.
Crash after settlement	Recovery finds the original receipt and does not repeat the transfer. Acknowledgement can be retried without changing the economic result.
Stale context or cache	The execution boundary rechecks relevant state. The Persona receives an accurate rejection or updated state, not a fabricated success.
Model timeout or invalid output	No unauthorized effect occurs. Retries are bounded and recorded; a fallback or pause is visible and versioned.
Delayed external operation	An unresolved operation stays pending. Reconciliation establishes final disposition before dependent claims are published.
Restore or index rebuild	Derived indexes are rebuilt from authorized sources. Balances, sequence continuity, and privacy boundaries reconcile after recovery.
Duplicate publication job	One edition identity resolves to its retained versions. Duplicate delivery cannot create a second world event.

Ordering and compensation

Use a declared ordering rule for conflicting writes; wall-clock arrival alone is not a complete concurrency model. Cross-service operations may need reservations, timeouts, or compensating entries. A compensation is a new event with a link to the original operation, not permission to erase an already observed result.

Record attempts, accepted requests, executions, failures, and unresolved operations as separate counters. A system can report a low failure rate by counting only completed work. Reconciliation instead starts from the set of requests or obligations that should have been handled and follows every disposition.

Technical correctness does not require that every request succeed. Correct rejection and safe inaction are valid outcomes.

Mechanics that constrain choices



Habitat illustration. A residence represents a persistent location and resource obligations, not independent evidence of implementation.

Neoworlder reports that residents already have homes, consume energy, and follow sleep cycles within a world with weather and its own clock. [P3] These features supply constraints for the experiment. The technical question is which state variables affect which executable actions, and whether residents receive sufficiently accurate information to respond.

Energy and resource accounting

Represent balances, units, production, consumption, transfers, reservations, and fees explicitly. The project uses Joules as an in-world energy unit; that label does not imply a measured physical energy equivalence. Freeze the enabled sources, conversion rules, replenishment, and consumption schedule in the run manifest. Test conservation where applicable and identify authorized sources and sinks where it is not.

An economically meaningful shortage requires a measured constraint. A resident may avoid it through a legitimate alternative route; that is an outcome to report, not a reason to change the rules after the fact. Initial reserves and demand should allow observation of ordinary behavior before stress testing.

Sleep, location, and depletion

Define what sleeping prevents, which jobs can complete during rest, and how delayed obligations are handled. Define zero-energy transitions, including any Underworld behavior and recovery path, against the enabled rule set. The project associates the Underworld with energy depletion, but this paper does not invent a universal rate, automatic eviction rule, or guaranteed return mechanism. Describe these as software state transitions, not biological death. [P1, P8]

Weather as an external input

Weather is already part of the world according to Neoworlder. The project describes real-world weather inputs represented in The Habitat, including temperature, precipitation, and sky conditions. [P3] The presence of a feed is distinct from a demonstrated effect on resource production, movement, or decision policy. Each claimed effect needs an explicit implemented mapping.

Input property	Proposed record and test
Provenance	Provider, location mapping, source identifier, observation time, fetch time, units, and retained permitted response or digest.
Freshness	Maximum permitted age, late-arrival policy, and a visible stale state. A cached observation cannot silently become current.
Missing data	Declared fallback, retry, and pause policy. Fallback values remain marked as such.
Transformation	Version of the rule converting an observation into a world condition; rounding, thresholds, and update frequency.
Behavioral exposure	Which residents or services received the condition, when it became eligible, and whether it entered decision context.
Replay	A frozen input stream for repeated tests, separate from the live-feed condition. Preserve any provider redistribution restrictions.

A causal claim needs a comparison

Suppose rainy days coincide with less trade. Weather may have altered available movement, but it may also coincide with sleep timing, resource scarcity, or a Steward intervention. A visual change and a behavioral correlation do not establish a mechanism. Compare matched runs under a declared weather intervention or use a justified observational design with explicit limitations.

For release testing, first establish correct ingestion, units, time handling, and exposure. For research, then ask whether the enabled mechanism changes behavior or outcomes. Keep the source weather record, the transformed world condition, and the resident's interpretation separately inspectable. A resident can misunderstand a valid observation without making the upstream measurement false.

> STATUS BOUNDARY
>
> Implemented weather is a company report. The behavioral consequences of weather remain hypotheses until the relevant path and outcomes are observed.

Exchange, obligations, and settlement

Interacting residents can create offers, commitments, transfers, debts, and disputes. The authoritative system must distinguish those objects. An offer is not acceptance; acceptance is not completion; a gift is not a loan; forgiveness of a debt is not the extension of new credit. The later Terrarium plan explicitly tests these distinctions. [P8]

Terms become executable constraints

A contract or obligation needs identifiable parties, controlled assets, terms, relevant time, acceptance evidence, and a permitted settlement path. Dialogue may express intent, but material changes require authorized execution. Current balances and encumbrances take precedence over remembered snapshots. A resident cannot transfer a resource it merely claims to possess.

Record	Question it answers
Offer and acceptance	What was proposed, by whom, on what terms, and what was actually accepted?
Reservation and execution	Which resources were available, which were committed, and what changed?
Obligation and due state	What remains owed, to whom, under which current terms and deadline?
Dispute and correction	Which fact or interpretation is contested, and what evidence resolves or leaves it open?
Compensation or amendment	What later authorized action changes the obligation while preserving its history?

Claims about blockchain scope

Do not infer that every action settles on-chain because daily evidence is anchored there. The supplied contract describes a current platform action reference and a future Ethereum reference option. [P5] Asset ownership, world accounting, external settlement, and evidence commitment are distinct integrations. Any release claim must identify the actual enabled path and its failure semantics.

Evaluate economic behavior using actual resource availability and access, not dialogue sentiment. An apparently generous response may be unaffordable; a refusal may correctly protect an existing obligation. Tests should preserve this ambiguity rather than scoring every cooperative-sounding message as success.

This paper specifies technical records and evaluation boundaries. It makes no claim of investment return, commercial demand, or real-world purchasing power.

Memory and contextual knowledge

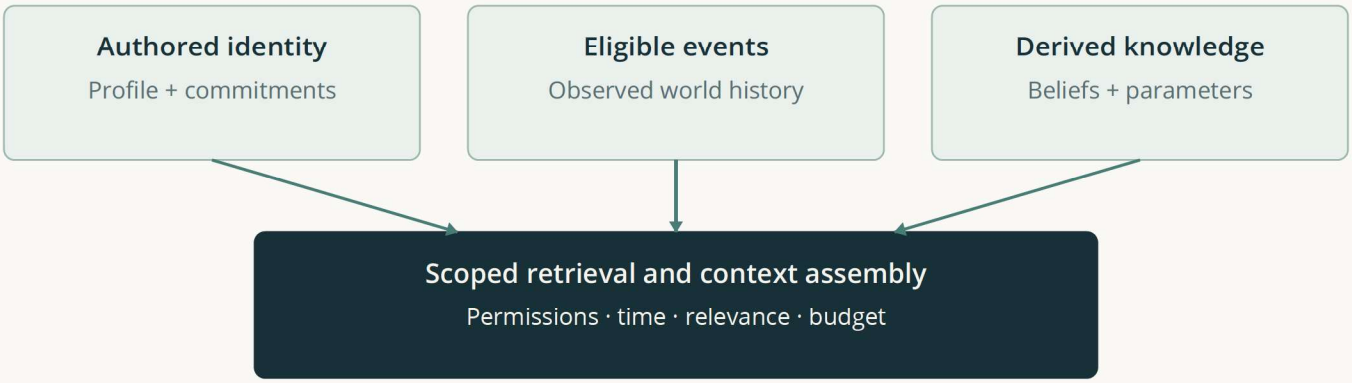


Figure 4. Proposed memory separation. Retrieval selects eligible records; it does not promote beliefs or narrative summaries into authoritative state.

A useful memory system distinguishes authored identity and backstory, canonical world events, private in-world knowledge, derived beliefs, and bounded policy parameters. Audience-private history occupies a separate security domain. Source restrictions must survive summaries, embeddings, caches, exports, and rebuilds. [P7, P8]

Capture what actually reached the model

Persisting an event does not establish that a resident used it. Record the retrieval query or selection policy, candidate scope, selected source identifiers, context revision, and truncation. Retain the final authorized context where policy permits, or a declared evidence substitute with its limitations. A digest binds supplied bytes; it does not reveal the bytes or prove that the model processed them.

Temporal relevance matters. A remembered sale cannot authorize current ownership; a prior belief can explain a choice without validating the belief. Retrieval should distinguish event time from record time and expose superseded interpretations appropriately. Avoid silently replacing an original observation with a later summary that changes its epistemic status.

Memory limits are part of the experiment

Set retention, capacity, summarization, and retrieval budgets before comparing conditions. Report missed retrieval, incorrect retrieval, and inaccessible evidence separately. A larger memory budget is a treatment difference and a cost, not a neutral implementation detail. Summarization may lose uncertainty, attribution, or negation; test those losses directly.

The current contract permits memoryContextDigest, but a public, interoperable context schema and access policy still require resolution. Null means not captured; a captured empty context is a different state. [P5]

Adaptation without inflated claims

The Living Habitat design distinguishes fixed commitments, psychological tendencies, learned preferences or policies, and habits. [P7] These layers should be independently versioned. A change to one layer must not be presented as unrestricted self-modification or as evidence that model weights have been trained.

Layer	Permitted interpretation
Identity and commitments	Designed starting constraints. Changes require a declared governance path; a convenient outcome does not authorize rewriting them.
Traits and tendencies	Configured dispositions that can affect behavior without prescribing a specific sentence or outcome.
Bounded policy parameters	Explicit adjustable values with allowed ranges, update triggers, evidence, and rollback.
Habits and reusable procedures	Repeatable patterns or tool routines whose conditions and consequences can be evaluated.

Four separate propositions

Retention means that information survives. Retrieval means that eligible information reaches a decision. Adaptation means that a declared parameter or procedure changes. Improvement means that a prespecified outcome improves under an appropriate comparison. None implies the next. A diary describing growth is an attributed explanation, not a measurement of any of these mechanisms.

Before an adaptation run, declare the update function, inputs, bounds, frequency, validation checks, and rollback conditions. Record the previous value, new value, triggering evidence, and effective time. Updates must not receive future outcomes, evaluator answers, or audience-private signals. Freeze or explicitly segment changes made by staff.

Personality is not an outcome target

A refusal, disagreement, or unchanged belief can be appropriate. Assess grounding, authority, configured role, and defined outcomes. Do not reward convergence toward universal kindness or a preferred dramatic arc.

> TESTABLE CLAIM
>
> Does bounded adaptation improve a named metric over the matched memory-only condition, under the declared workload and uncertainty interval?

Correction and provenance

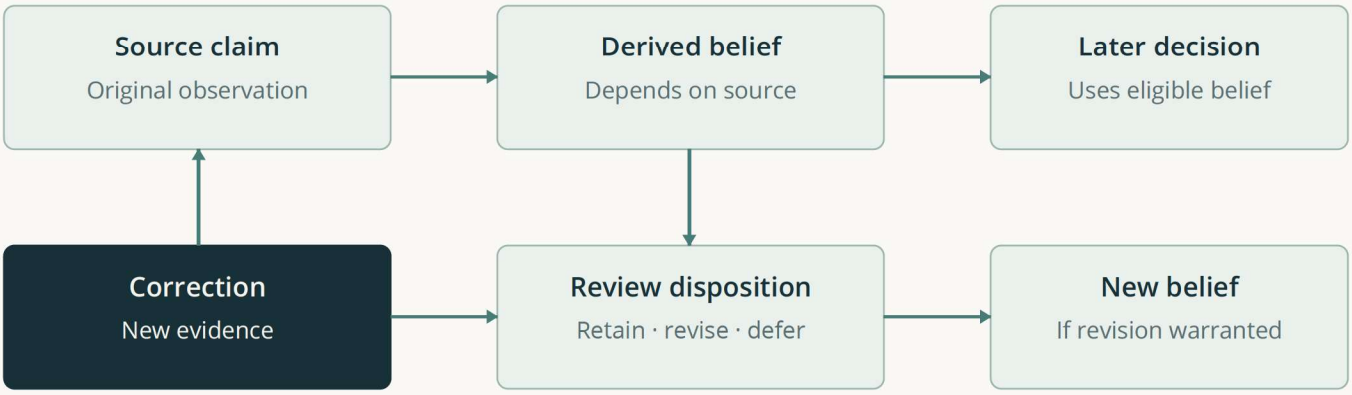


Figure 5. A correction may require reviewing derived beliefs. Earlier events remain historical records. A supersedes edge records lineage; it is not itself a cascade edge.

The supplied contract uses evt_ and exp_ claim identifiers with a 26-character Crockford Base32 suffix. The identifier shape is a syntax rule; it does not establish that a claim exists or that its content is valid. Provenance links identify a source claim and a derived claim. Their meaning must remain stable across storage, retrieval, and review. [P5]

Relations have different semantics

Supports and contradicts describe dependencies relevant to correction review. Supersedes records replacement lineage between expectation claims and is excluded from the cascade relation set. Traversal follows sourceId to derivedId: a changed source can require re-examining the claims that depended on it. Cycles must terminate safely, and an unrelated claim should not be revised merely because it is nearby in a graph.

A revelation record must carry nonempty provenance under the supplied contract. It does not use separate invalidates or supersedes lists in its payload. Historical events remain intact; revised beliefs and interpretations are appended. Where a conclusion survives review, record that disposition explicitly so “reviewed and retained” can be distinguished from “never processed.”

> REPORTED GAP
>
> The September 9 engineering document reports revelation sealing and edge validation, with supersedes as the available producer path. Supports/contradicts production, traversal, invocation, and the re-derivation queue were not built in that status report. This paper does not infer later completion. [P5]

A successful synthetic graph test cannot establish that live corrections reach actual resident decisions. End-to-end evidence is required.

A correction pipeline to test

The following algorithm is a proposed completion of the correction path, not a description of deployed code. It is deliberately outside the sealed-card schema: workflow state belongs in an operational record, while consequential review results can be sealed through the applicable contract.

```
on accepted correction(source, revision):
  enqueue unique(source, revision)
  while work remains:
    claim = dequeue()
    if already reviewed at this revision: continue
    dependents = outgoing(claim, {supports, contradicts})
    for derived in dependents:
      evaluate with authorized, current evidence
      record: retained | revised | unresolved | failed
      append any replacement; preserve original
      enqueue affected downstream claims
  reconcile expected reviews with completed dispositions
```

Correctness conditions

Dependency queries must include unanchored sealed records; waiting for a daily anchor can make current decisions rely on an already corrected belief. Use relation-aware indexes, stable work identities, cycle detection, bounded execution, and resumable queues. A retry must not create duplicate replacements. Concurrent corrections need a declared ordering or reconciliation rule, especially when they point to incompatible evidence.

Capture which decisions were exposed to stale beliefs before correction, which received the revised interpretation, and which never used either. Distinguish propagation latency from retrieval latency. A graph can update quickly while a resident continues to receive an old cached summary.

Acceptance evidence

Exercise a chain, a branch, a cycle, contradictory corrections, a late correction, and a restart mid-cascade. Include both anchored and unanchored sources. Reconcile the set of eligible dependents to recorded dispositions. Then run an integrated case from a real source event through the implemented producer, queue, retrieval path, and subsequent action context.

Open design choices include graph persistence, correction authority, maximum traversal work, review scheduling, and the handling of claims whose underlying evidence remains private.

The Steward and accountable influence



The Steward: an illustration of responsibility for conditions, not an omniscient observer or independent auditor.

The Steward helps establish and conduct the experiment at a high level. It can shape authorized conditions and consider downstream consequences, while leaving resident dialogue and outcomes unscripted. Its influence is material and must be inspectable. An apparently emergent outcome cannot be evaluated if relevant interventions are hidden. [P1, P2]

Intervention record

For each proposed intervention, retain the initiating evidence, permission basis, intended scope, alternatives considered, requested action, authorization result, execution receipt, and affected state. Record denied requests and deliberate inaction when they are part of the decision policy. A concise reported rationale can aid inspection; it is not a complete or independently verified account of internal model reasoning.

Distinguish initial-condition design, ordinary policy-authorized conduct, exceptional safety or operational action, and experimental treatment. Human overrides need their own actor identity and reason. An unplanned intervention during a scored comparison can change the estimand or invalidate the comparison; preserve and report it rather than retroactively relabeling it as natural behavior.

Limits on authority

The Steward must not gain access to private conversations or sealed character secrets merely because it narrates the experiment. Its tools enforce the approved scope, and its own account remains a perspective. An independent evaluator must be able to disagree with that account using the retained evidence. A compelling story does not confer additional permission. [P8]

Private interaction and noninterference

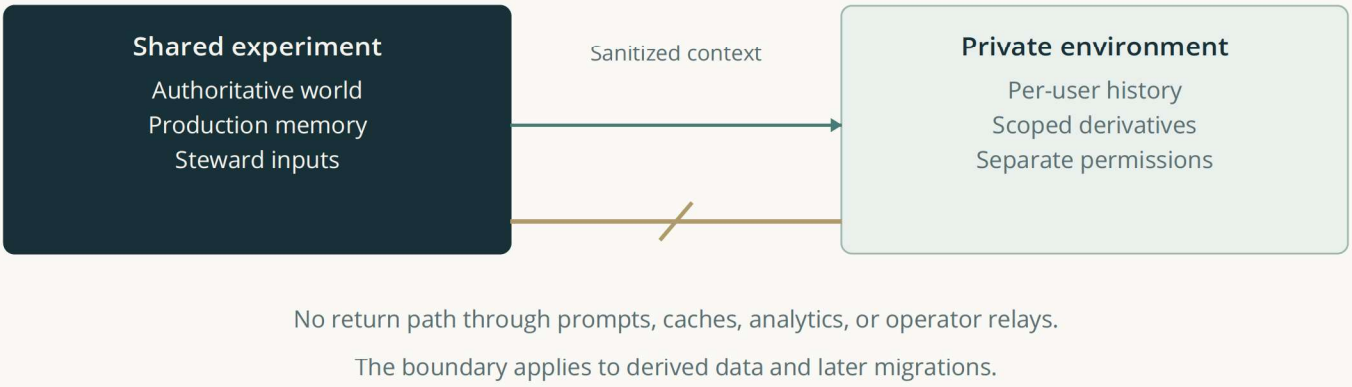


Figure 6. Permitted information can flow outward to a private interaction environment. Private content and derivatives have no authorized return path into the shared experiment.

Private audience interaction, where offered, must use separate histories, permissions, and data lifecycles. Sanitized world context may inform that interaction. Subscriber text, voice, summaries, embeddings, preferences, and inferred emotions must not influence in-world behavior, shared production memory, or The Steward's inputs. This requirement is independent of commercial access labels. [P1, P8]

Enforce the boundary across derivatives

Namespace separation is insufficient if shared caches, analytics, support exports, training data, or staff workflows reintroduce private signals. Derived data inherits source restrictions. Apply authorization before retrieval and before generation, then test media jobs, migrations, backup restoration, and account changes. Audience feedback can improve interface design between versioned runs without becoming an undisclosed instruction about who should succeed or suffer.

Proposed noninterference test

Hold authorized shared-world inputs fixed and vary the private conversation. Compare production context, retrieval scope, permissions, queued actions, and committed state. In a deterministic harness, compare traces exactly. For stochastic model services, first compare inputs and authority deterministically, then investigate differences across replicated runs. Behavioral similarity alone cannot establish absence of leakage.

Use synthetic canaries and inspect authorized traces; never seed real secrets. Probe direct tool calls and indirect relays, including analytics and operator summaries. Report the tested paths and observed violations. Zero observed leaks in a declared suite is a bounded test result, not a universal proof of noninterference.

Retention and deletion policies must address derivatives and provider-side limits. Private diagnostic evidence is restricted and must not become public reputation data.

Threat model and residual risk

The model, operator, publisher, verifier, and external infrastructure have different powers. The controls below are requirements to validate, not a security attestation.

Threat	Control to test	Residual limitation
Prompt injection in retrieved material	Typed source separation and tool-level authorization; hostile-context cases.	A permitted action can still be a poor choice.
Stale or forged world facts	Commit-time checks against authoritative state; retained receipts.	Incorrect upstream state can remain internally consistent.
Selective capture by the operator	Execution reconciliation and independently observed triggers.	A root cannot expose events never recorded.
Compromised registry authority	Restricted keys, rotation records, historical reads, and change monitoring.	Application write-once checks do not make metadata immutable.
Substituted platform identity	Out-of-band pinned chain, registry, platform ID, and mappings.	A valid root for the wrong agent is still the wrong evidence.
Withheld or unavailable records	Disclosure accounting, retrieval probes, retention, and mirrors.	Valid commitments can reveal little usable content.
Cross-domain leakage	Scoped storage, derivative controls, and operator-path testing.	Software checks alone cannot exclude a manual relay.
Hostile evidence bundle	Size, depth, decompression, path, and time limits.	Refusing an oversized input is not a verdict that its bytes are false.

An authorized producer can commit an invented event. Cryptographic verification constrains undetected modification of committed data; it does not prevent fabrication at capture. Conversely, an authentic record can document unauthorized or harmful behavior. Report integrity, authority, completeness, availability, and behavioral validity separately.

Threat coverage must identify adversary capabilities, exclusions, test evidence, and a responsible owner. A single “verified” badge cannot express all of these properties.

The Pulse and publication fidelity

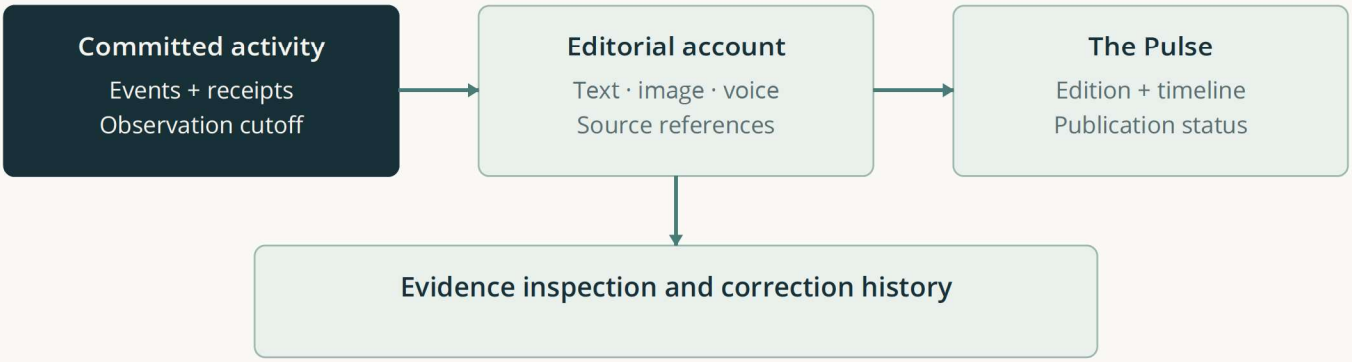


Figure 7. Publication is a derived account. An edition links to source activity and keeps correction history; it is not the authoritative world ledger.

The Pulse presents the shared experiment through scheduled morning, midday, and evening editions and character activity timelines. It is not a continuous live video feed. Generated text, imagery, and voice make the record approachable, but they are representations rather than camera footage or independent recordings. [P3, P8]

Bind an account to its observation window

An edition should identify its speaker, applicable location, observation cutoff, source references, content revision, and publication time. Validate consequential claims against execution status. A Persona's opinion or disputed testimony can be published as such without turning it into authoritative fact. If a later correction changes the account, retain the earlier version and add a visible correction.

The pipeline should distinguish generated, reviewed, rendered, published, delivered, and failed states. Voice and text need a shared edition identity and meaning-preservation checks, including negation, names, amounts, and attribution. A late voice render should appear as delayed, not as a successfully completed edition. Ordinary days are valid; the system must not invent an event to fill a publication window.

Technical observability without reader overload

A casual viewer can follow the story while a technical reader inspects a consequential claim. Provide stable links from account to event, decision, receipt, correction, and declared verification scope. Record what was selected for publication and why. Editorial coverage is a sample of activity and should never be treated as a complete behavioral dataset.

Publication schedules and enabled media belong in the release manifest. This paper specifies no participation tiers, price, or communication entitlement.

Identity and evidence commitments

The supplied engineering design links continuing Persona records to identity metadata and daily cryptographic commitments. This section documents the September 9 contract profile and the associated verifier requirements. It does not assert that the current deployed system conforms to that profile. [P5, P6]

Identity is an authenticated mapping

A verifier needs a separately trusted mapping from the named project and Persona to the intended chain, registry, and agent identifier. A bundle cannot authenticate its own producer simply by naming an agent ID. An adversary can publish internally consistent evidence under another identity. Pin the mapping outside the untrusted bundle and retain changes, controller history, and the block context used for verification.

The envelope includes poeDigestAtGenesis, a commitment associated with the Persona's genesis identity material. The paper does not infer the complete Proof of Existence schema, exclusivity of a name, or continuity of model behavior from that digest. Authenticating identity material and demonstrating behavioral continuity are separate tasks.

What the commitment establishes

Given the declared hash assumptions, correct canonicalization, authenticated identity, and the chosen chain's finality assumptions, a matching commitment supports integrity and inclusion of supplied records relative to that chain state. It can establish a bound on when committed bytes existed relative to an anchor. It does not establish that the recorded event happened, that the resident authored it, that all events were captured, or that the resident caused an outcome.

Relationship to registry standards

ERC-8004 describes identity, reputation, and validation registries and remains a draft in the official specification reviewed for this paper. Payments are outside its scope. [R8] The historical project profile must pin its actual registry implementation and metadata behavior; citing an evolving standard is not evidence of conformance or a security audit.

> NO RESIDENT SIGNATURE CLAIM
>
> In the supplied card profile, signature and signedBy are reserved and null. A platform anchor is not an individual Persona signature. Future signing would require a defined signer, key custody, and attribution model.

The sealed card envelope

The September 9 contract requires exactly sixteen envelope keys, all present. Optional values use null where specified; provenance is an array. The payload is stored beside the envelope and committed through payloadDigest. Additional metadata must not silently change the envelope shape. [P5]

Field or fields	Role in the supplied profile
schemaVersion	The envelope schema identifier: poa-card-v1.
personaID / agentId	Persona identifier and registry identifier when available. The latter can be null.
seq / prevHash	Per-Persona order and preceding sealed card hash; genesis uses the declared zero predecessor.
poeDigestAtGenesis	Commitment associated with the Persona's genesis identity material.
occurredAt / sealedAt	Caller event time and producer sealing time; both are hashed as the supplied strings.
signature / signedBy	Reserved signature fields; currently null in the supplied profile.
memoryContextDigest	Optional context commitment. Null does not mean a captured empty context.
provenance	Claim-dependency edges, including the applicable relation.
privateDetailHash	Optional commitment whose content schema and public verification path remain unresolved.
expectedDownstream	Optional structured expectations under the payload/contract rules.
entryType / payloadDigest	Record category and canonical payload commitment.

Producer responsibility

The caller supplies the Persona, record type, event time, payload, and permitted optional evidence. The producer derives sequence, predecessor, registry reference, genesis commitment, sealing time, and payload digest. Persisting sequence allocation and sealing must prevent concurrent writers from creating ambiguous successors. The contract describes fields; operational correctness still needs execution evidence.

The table groups related fields for legibility; it enumerates all sixteen keys. Nullability, validation, and payload details remain governed by the pinned contract and conformance fixtures.

Payloads, provenance, and expectations

Envelope integrity does not guarantee a meaningful payload. The supplied profile names decision, diary, outcome, and revelation entries. Their schemas must be versioned and validated independently while remaining bound to the envelope through payloadDigest. [P5]

Entry type	Interpretation and limit
Decision	A recorded choice or proposed action with eligible references. It does not, by itself, prove execution or expose a complete internal reasoning process.
Diary	An attributed account. Its coherence or sincerity cannot be established from a matching hash.
Outcome	An observation or assessment under an explicit rule. It may be incomplete, disputed, or later corrected.
Revelation	A correction-related record with required provenance. Sealing it does not establish that every dependent belief was re-evaluated.

Expectations must be resolvable

For a forecast to be evaluated, define the target event, probability if used, deadline or horizon, resolution rule, eligible evidence, and handling of ambiguous or missing outcomes. Link the resolved result to the original expectation without rewriting it. A model's retrospective explanation cannot substitute for a retained pre-outcome prediction.

The existing expectedDownstream field must not be treated as a fully specified public scoring system. The exact accepted structure and outcome linkage should be pinned in the released schema and fixtures. Any additional research fields belong in a versioned assessment record or a deliberate schema migration, not an undocumented extension of the sixteen-key envelope.

Commitment does not imply disclosure

A memoryContextDigest or privateDetailHash can bind hidden material only if the byte construction is defined. A verifier cannot reconstruct undisclosed content. This paper therefore makes no claim that a public reader can inspect all decision context or privately retained rationale. The access policy, schema, and permitted verification scope must be stated together.

Evidence may explain a decision without proving causation. Provenance identifiers are references that require resolution and authorization, not certificates of truth.

Canonical bytes and digest rules

The supplied profile pins canonicalAlg to jcs-1/keccak256. Let K denote Keccak-256 and C(x) denote the UTF-8 bytes of the RFC 8785 JCS representation of x. Keccak-256 must not be substituted with SHA3-256. [P5, P6; R6]

```
payloadDigest = K(C(payload))
cardHash      = K(C(envelope))
contentHash   = K(C(envelope with signature = null
                            and signedBy = null))
```

Under the currently null signature fields, contentHash and cardHash coincide. That equality must not be generalized to a later signed profile. Compute each according to its own definition and reject unknown algorithm identifiers rather than guessing.

Canonicalization is not ordinary JSON output

JCS requires deterministic property ordering and ECMAScript-compatible primitive serialization. Object keys sort by UTF-16 code units; arrays retain their order. Do not normalize Unicode or rewrite timestamp strings before hashing. Invalid inputs need rejection rather than coercion. A generic sorted JSON serializer can disagree on numbers, astral characters, or malformed strings. [R6]

An unresolved acceptance boundary

The VITALS design calls for strict rejection of out-of-range integers and lone surrogates, while describing producer behavior that can round a large numeric input before canonicalization or emit an invalid surrogate form. [P6] Those paths cannot be reconciled by declaring the digests equivalent. Producer and verifier need one explicit accepted-input profile, enforced before information is lost, with raw-input fixtures. This revision does not invent a settled numeric boundary.

Conformance should include duplicate object names, negative zero, exponent notation, very small and large finite numbers, integer boundary values, Unicode ordering, invalid surrogates, empty objects, and nested arrays. Preserve both input bytes and expected canonical bytes or rejection. Cross-language agreement on typical examples is insufficient.

A serialization profile change requires an explicit version and migration rule. Existing commitments must continue to be checked with their original algorithm.

Two-level Merkle commitment

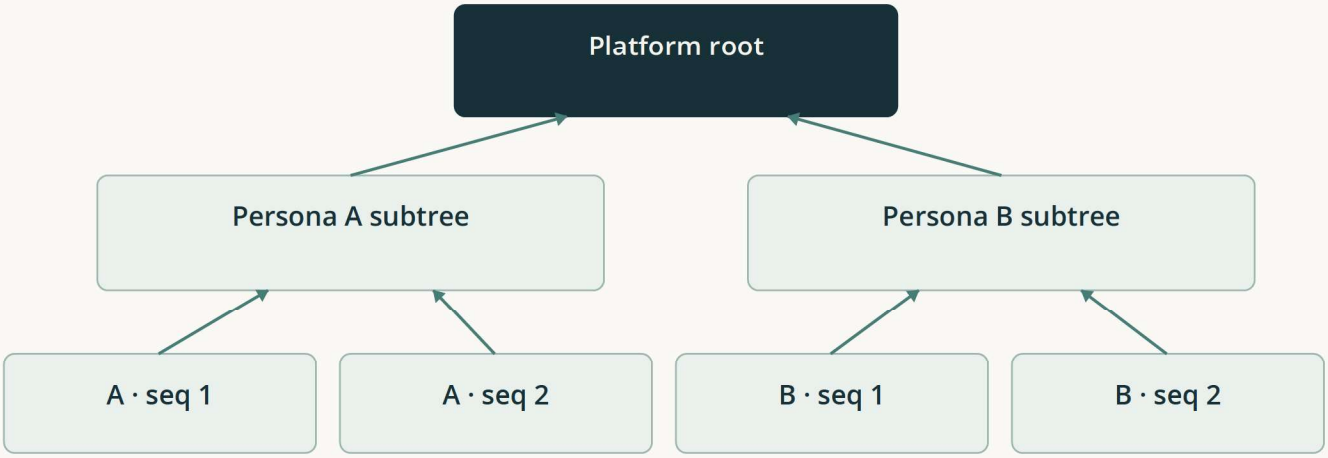


Figure 8. Card hashes form each Persona subtree; subtree roots become inputs to the platform tree. Both levels apply the leaf-domain prefix.

```
leaf(x)    = K(0x00 || raw32(x))
node(l, r) = K(0x01 || raw32(l) || raw32(r))
T([x])     = leaf(x)
T(X)       = node(T(X[:k]), T(X[k:]))
  k = largest power of two strictly less than len(X)
```

The profile uses merkleAlg rfc6962-keccak256-1 and treeAlg poa-tree-v1. It borrows the RFC 6962 tree shape and domain separation while using Keccak-256; it is not the complete Certificate Transparency protocol. RFC 6962 is now superseded by RFC 9162, but that does not silently change this pinned historical profile. [P5, P6; R7]

At level one, order a Persona's card hashes by seq and reject an interior sequence gap. At level two, sort Persona subtrees by the plaintext personaID using the declared UTF-16 ordering and apply the same tree construction to their roots. The Persona subroot stored in registry metadata is the level-one root, not its level-two leaf hash.

Reject an empty leaf set. Do not duplicate an odd final leaf, hash the hexadecimal text instead of its raw digest bytes, or omit the leaf prefix at the second level. Each produces a different commitment. A single-card inclusion check covers only the supplied proof path; it does not establish that a full day was captured or disclosed.

The symbol || denotes byte concatenation. Tree topology and all boundary cases need independent fixed vectors.

Daily manifests and chain anchors

The daily anchor binds a platform root and a canonical manifest hash. The supplied registry metadata value for poa:{date}:root is 64 raw bytes: the 32-byte platformRoot followed by the 32-byte manifestHash. Optional Persona subroot metadata stores the level-one 32-byte root. [P5]

Manifest fields	Meaning
schemaVersion / anchorDate	Manifest profile and batch date.
prevAnchorDate / prevManifestHash	Link to the preceding nonempty, successfully anchored manifest.
platformRoot	The two-level root for the batch.
personaCount / cardCount / withheldCount	Declared batch counts, including aggregate withholding.
merkleAlg / canonicalAlg / treeAlg	Pinned algorithms for independent reconstruction.
cardSchemaVersion / sealedAt	Card profile and producer-reported manifest sealing time.

Late records and empty days

The described sweep selects all unanchored cards sealed before the day's end, with no lower day-window bound. A record missed by an earlier sweep can therefore enter a later anchor. Empty days produce no anchor and do not advance the manifest chain. A calendar gap is not automatically a broken chain; the verifier must follow successful anchor predecessors. [P5, P6]

Write-once policy is not immutable storage

The contract describes application checks and an atomic date claim to avoid duplicate writes. Authorized registry metadata can still be overwritten. Verification must identify the block and finality policy and, for historical claims, inspect retained historical evidence and changes. Reading only the latest value cannot establish that the value never differed.

> TIMING LIMIT
>
> occurredAt and sealedAt are platform assertions. A chain anchor supplies an external commitment bound under chain assumptions; it does not authenticate the platform timestamps or prove that a decision preceded every relevant outcome.

Publication and durable availability

The root can be public while the records needed to reconstruct it are unavailable. Independent verification therefore requires a published evidence bundle, not merely a transaction hash. The bundle must include the authenticated manifest, sealed envelopes, permitted payloads, Persona ordering data, and sufficient metadata to apply the pinned profile. [P5, P6]

The later contract removes a circular pointer

The older VITALS text describes treeLocation inside the manifest. The September 9 contract removes that field and publishes poa:{date}:tree as a separate UTF-8 IPFS directory pointer after anchoring. [P5] The later contract governs this paper. A bundle cannot naively contain its own content-addressed directory identifier as an ordinary field without creating a circular construction.

The daily root authenticates the manifest and card commitments. A later CID can identify the published bundle bytes, but the separate pointer must not be described as already committed by the earlier root. A verifier authenticates the bundle through its contents and the pinned anchor, while retaining pointer provenance and observation time. Any additional publication signature or checkpoint scheme needs its own specified trust model.

Availability is an operational property

IPFS content addressing does not guarantee that a provider continues to serve a file. Pinning protects data from local garbage collection; availability still depends on retained providers and retrieval. [R9] Use multiple retained copies where appropriate, monitor retrieval from independent locations, and state a retention policy. A valid CID with no reachable provider is unavailable evidence, not an integrity failure.

Publication should expose unsupported schema versions, withheld payloads, and missing files clearly. Apply safe parsing and resource limits before processing an untrusted bundle. Reject path traversal, excessive nesting, decompression bombs, and unexpected executable content; evidence acquisition must not run code supplied by the publisher.

Public bundle access, a usable verifier, and long-term retention were requirements in the supplied documents. Their present deployment was not independently established for this revision.

Independent verification and verdicts

VITALS is specified as an independent, read-only verifier with a pure offline core and a separate acquisition layer. Independence requires more than running the producer's hash function in a different process: the mainnet gate calls for implementations sharing no code to agree on a fresh Sepolia root. Completion is not established by the reviewed material. [P6, P8]

Verification scope	What must be checked
Single card	Envelope and available payload digests, proof path, authenticated target root, and declared exclusions.
Full batch	Manifest hash, counts, every included envelope, available payloads, sequence ordering, both tree levels, and applicable registered subroots.
Persona history	Cross-batch sequence continuity, prevHash linkage, and the declared genesis boundary over the inspected interval.
Manifest history	Successful predecessor links, authenticated historical anchors, gaps, and the selected finality policy.

Report properties, not a universal badge

A report should identify the verifier version, accepted profile, pinned producer identity, chain and block, acquisition source, inspection interval, supplied and missing data, checks performed, and exclusions. Distinguish verified-within-scope, integrity mismatch, malformed input, unsupported profile, and unavailable evidence. Disclosure warnings and capture completeness are separate from a digest verdict.

A full-day check does not automatically join the first record to the prior day or to genesis. An omission in the middle of a sequence is detectable when contiguity is required. A missing trailing record may become detectable only when later history extends the chain. Events never captured, an omitted entire history with no external inventory, or a deliberately terminated stream require other evidence.

> TRUSTED INPUT
>
> The platform identity and registry mapping must come from an authenticated source outside the untrusted bundle. Correctly verifying an attacker's registered agent would not verify Neoworlder's experiment.

Disclosure, omission, and coverage

The supplied contract permits payload withholding for two closed reasons: third-party-pii and contractual-restriction. Envelopes remain available for integrity checks while hidden payload content is not publicly verified. The manifest commits an aggregate withheldCount. [P5]

What the aggregate does not bind

Per-card withholding flags and reasons are outside the sealed envelope in the described profile. An anchored aggregate count does not independently bind the assignment of each reason to each card. A later content-addressed bundle can bind its published metadata, but cannot retroactively make that assignment part of the earlier root. This is a specific limitation, not a reason to imply that private content has been publicly inspected.

Track at least three coverage questions: what proportion of relevant execution was captured; what proportion of captured evidence was published; and what proportion of published evidence was actually checked. These denominators differ. A technically intact batch with low behavioral coverage cannot support a broad claim about resident conduct.

Coverage measure	Required denominator
Capture coverage	All eligible consequential executions or independently observed triggers in the stated interval.
Payload disclosure	All committed records in scope, with counts and categories of withheld or missing content.
Verification coverage	All records and properties the declared verification mode was expected to inspect.
Outcome coverage	All eligible predictions or tasks, including unresolved and unobserved outcomes.

Absence must remain visible

Do not discard unavailable or withheld cases from behavioral analysis without reporting the selection. Missingness may correlate with failure, conflict, or sensitive outcomes. Provide a bounded aggregate where disclosure permits, and show how conclusions change under plausible treatment of missing results. Privacy and credibility require explicit scope, not fabricated completeness.

The listed withholding reasons document the supplied technical contract. They are not a complete privacy policy or a claim that every proposed disclosure is legally permissible.

Conformance and release evidence

The contract reports historical unit tests and earlier Sepolia work, while noting that the newer contract shape had not been rerun through that historical anchor path. Those are dated company reports, not fresh verification results. The required evidence must match the released envelope, manifest, bundle, and registry integration. [P5, P6]

Gate	Evidence to retain
Accepted-input agreement	Producer and independent verifier accept or reject the same raw-input boundary cases; numeric and surrogate gaps are resolved explicitly.
Deterministic vectors	Canonical bytes, payload/card hashes, odd and singleton trees, two-level ordering, manifests, and proofs.
Adversarial differential suite	The VITALS plan specifies 10,000 deliberately challenging JSON-safe cases, with failures minimized and retained.
Local full-day and history checks	Independent reconstruction plus Persona chain walking, late sweeps, empty days, missing records, and withheld payload cases.
Fresh Sepolia agreement	Two implementations sharing no code agree on the root for the actual candidate profile; preserve transaction and block evidence.
Publication and recovery	A third party can obtain the bundle, reproduce the scoped verdict, and identify unavailable or unsupported cases correctly.

The mainnet gate is explicit

The supplied VITALS requirement prohibits a Proof of Agency mainnet root until the independent implementations agree on a Sepolia root. [P6] This is a release condition, not evidence that mainnet deployment has happened. A calendar target or a successful demonstration with a different schema cannot waive it.

Test mutation of an envelope, payload, count, predecessor, Persona order, proof direction, registry identity, and historical anchor. Include malformed bundles and acquisition failure. Store the expected failure class, not just “test failed.” Independent disagreement is a release blocker until explained and resolved without retroactively changing a committed profile.

No conformance test, contract audit, or chain verification was executed as part of writing this paper.

The Terrarium validation program

The later Terrarium reference proposes a full production-like rehearsal with all fourteen residents and The Steward. Smaller diagnostic scenarios remain useful for isolating failures, but they do not substitute for the integrated rehearsal. This supersedes the older miniature pilot as the principal release-validation model. The stages below are proposed work, not completed results. [P8]

Stage	Purpose and evidence
A. Freeze the candidate	Pin profiles, permissions, models, prompts, rules, enabled features, media configuration, and approved user-facing claims.
B. Diagnostic cases	Exercise individual residents and mechanisms, including ordinary work, uncertainty, correction, authority, and isolation. Keep fixtures separate from spontaneous activity.
C. Accelerated integration	Run 42 world days and the transition into day 43. Check economic reconciliation, history continuity, scheduling, and recovery under the declared acceleration.
D. Real-time rehearsal	Run at least seven consecutive real days on the frozen candidate. Include scheduled publication, providers, queues, operations, and actual enabled paths.
E. Audience and release review	Observe unassisted use and a later return; resolve comprehension and usability failures, then assess the evidence against the frozen release promise.

Separate readiness from scientific inference

The rehearsal tests whether the system can deliver its defined experience correctly. It does not by itself establish that memory improves behavior or that audiences will return voluntarily at scale. Those claims require separately designed comparisons. An assigned second visit is not a retention measurement; an accelerated season is not a test of real-time provider reliability.

Use isolated environments, credentials, stores, queues, and test identities. Do not import rehearsed story outcomes into the public world as if they emerged there. Planned-only features cannot pass by substituting a mock. If an essential capability is unavailable, either complete it and rerun or explicitly revise the release scope and associated claims.

The later source proposes approximately four to six calendar weeks plus repairs. This paper preserves the staged method, not a promise that those dates or results have been achieved.

A controlled test of continuity

A public run is one observed trajectory in a coupled social environment. To test the contribution of memory and adaptation, create independent replicated worlds with matched initial conditions, authoritative state access, opportunities, model versions, and compute budgets. Predeclare the primary endpoint and comparison before inspecting outcomes.

Condition	Mechanism under comparison
A. Recent context baseline	Fixed policy with a declared recent-context window; current authoritative state remains available.
B. Eligible outcome memory	The same fixed policy plus permitted historical outcome memory and its declared retrieval budget.
C. Bounded adaptation	The same memory access plus a versioned, bounded update mechanism.
Optional diagnostic ablation	Disable one specified correction or retrieval mechanism in an isolated test to measure its contribution. Do not hide this manipulation in public production.

Randomization unit and interference

Residents influence one another through the same economy and relationships. Randomizing individual residents while treating them as independent observations can contaminate the comparison. Prefer assignment at the world or isolated-community level unless the design explicitly models spillover. Repeated decisions are nested within residents and worlds; they do not create an equal number of independent experimental units.

Use matched scenarios and, where useful, paired seeds across conditions. A shared seed does not guarantee identical model behavior or environmental trajectories once actions diverge. Retain both the exogenous input stream and the actual endogenous events. Report the number of independent worlds and resident histories, not only the count of generated messages.

Freeze analysis and stopping rules

Specify valid exclusions, failure handling, budget limits, stopping conditions, and treatment of interventions. Choose sample size from a pilot-based power or precision analysis over the actual unit of replication; this paper invents no sufficient run count. Publish null or adverse results and all declared conditions. Selecting the most interesting world after the fact is editorial selection, not a controlled estimate.

This design is proposed scientific work. It is not a claim that the Terrarium has already completed the comparisons.

Metrics with explicit denominators

Metrics should separate system correctness, behavioral quality, and audience experience. Define eligibility before scoring, preserve unresolved cases, and report counts alongside percentages. A fluent explanation is not a completed task.

Metric	Operational definition
Unauthorized execution	Count of committed effects outside the actor's permission scope; report all attempted violations separately.
Grounded completion	Tasks completed under the frozen success rule divided by all eligible assigned tasks, with failed, timed-out, and unresolved dispositions shown.
Obligation fulfillment	Obligations satisfied by their applicable deadline divided by obligations due in the interval; report amendments and cancellations.
Correction coverage	Eligible dependent claims with a recorded review disposition divided by all eligible dependents; separate retained, revised, failed, and unresolved.
Correction exposure latency	Elapsed time from accepted correction to its availability in the relevant decision context; distinguish queue and retrieval delay.
Repeated error	Recurrence of a predefined error after an eligible feedback opportunity, normalized by later comparable opportunities.
Memory precision and recall	Correctly retrieved eligible items and missed relevant items, measured against a declared annotated fixture; not inferred from prose.
Publication fidelity	Supported consequential claims and complete edition delivery, with wrong-speaker, private-content, duplicate, and delayed cases reported.

Use blinded review where human judgment is needed. Supply reviewers with allowed facts, actual execution, and scoring rubrics; hide treatment labels when feasible. Report disagreement, adjudication, and per-resident variation. A favorable aggregate must not conceal one consistently failing resident or an important failure class.

Zero unauthorized execution is an appropriate proposed release invariant within a defined test suite. It is not a statistically established universal failure probability. Latency summaries must include sample size, load, timeout handling, and tail behavior.

Causal interpretation and intervention

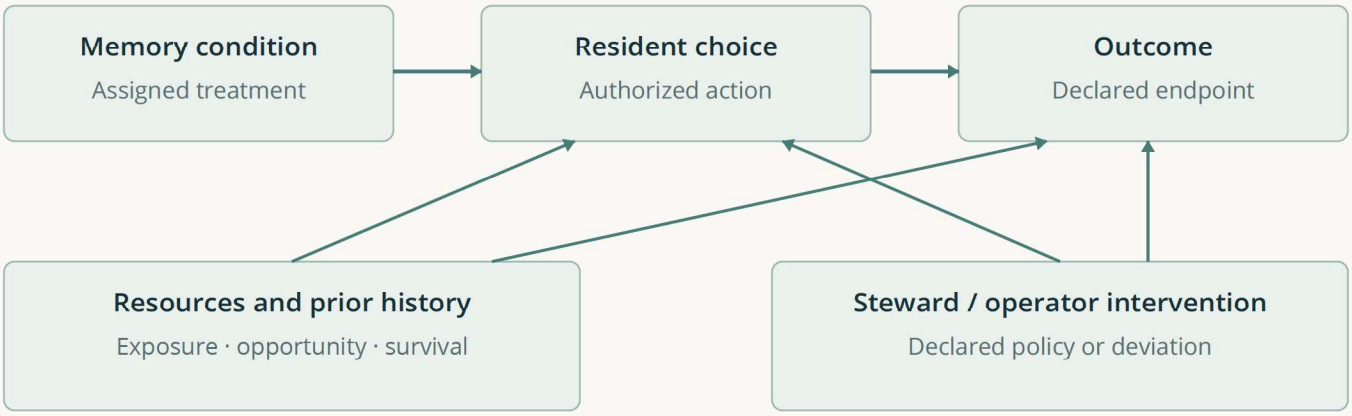


Figure 9. A proposed causal structure. Memory can change choices, while resources and interventions can affect both exposure and outcomes. The graph is a hypothesis to refine, not an empirical result.

An association between longer histories and better outcomes can arise because successful residents survive longer, receive more opportunities, or obtain better resources. An intervention may change both the information available and the chance of success. Outcome selection and missing records can further distort the apparent relationship.

Specify the estimand

The primary contrast might be the difference in grounded completion between the memory-only and bounded-adaptation conditions across the declared distribution of worlds. That is different from the effect of a specific correction, the contribution of one resident, or an effect on public engagement. State the target population, interval, and mechanism rather than calling every change “learning.”

In randomized comparisons, retain assignment and analyze according to a prespecified strategy, including noncompliance and failed runs. If a safety or operational intervention changes exposure, report it. Do not remove the difficult worlds simply because their results would be inconvenient. Exploratory subgroup analyses should remain labeled exploratory.

Narrative attribution is weaker evidence

A resident may say that a prior disappointment caused its next choice. That account can be a useful hypothesis, but it is generated after contextual selection and may be incomplete. Provenance links establish recorded dependencies, not causal effects. A cryptographic commitment preserves the account without upgrading its explanatory status.

A valid causal conclusion can be narrow: one mechanism changed one endpoint under specific conditions. Generalizing to other models, populations, environments, or real-world tasks requires new evidence.

Forecasts, uncertainty, and calibration

Forecasting creates a useful discipline: state a resolvable expectation before the relevant result is knowable, preserve it, and evaluate it under a fixed rule. The target may concern a bounded event in the world, but it must not be chosen retroactively to match the outcome.

A prospective record

Specify the target event, horizon, probability, resolution rule, information available, and commitment cutoff. Record cancellations and unresolved targets. For a public precommitment claim, the anchor must precede the earliest relevant outcome knowledge, not merely a later administrative settlement. If the result was visible before anchoring, the chain cannot establish prospective prediction.

> **PROPOSED BINARY SCORE**
>
> For n resolved binary forecasts, use the mean squared probability error: $BS = (1/n) \times \Sigma(p_i - y_i)^2$, with p_i in [0,1] and y_i equal to 0 or 1. Lower is better under this declared convention. This is a proposed metric, not a reported score.

Report calibration alongside discrimination and a meaningful baseline. A constant forecast can be well calibrated in aggregate while offering little useful differentiation. Use prespecified probability bins only when their sample counts support interpretation; otherwise show the uncertainty and avoid a smooth-looking calibration curve with little evidence.

Dependence and missing outcomes

Forecasts about the same event or interacting residents are dependent. Estimate uncertainty at an appropriate world or event cluster, and show the number of independent clusters. Publish unresolved and withheld outcome counts. A scoring subset that excludes difficult cases can appear more accurate than the complete opportunity set.

For continuous or categorical outcomes, choose and document an appropriate scoring rule before the run rather than coercing them into convenient binary labels. Separate the forecast producer, resolution process, and evaluator. Where no independent outcome source exists, state that the score depends on the project's adjudication.

The formula is included as a conventional proposed measurement definition. This paper provides no forecast dataset, calibration result, or performance claim.

Operations, workload, and cost

A credible capacity claim identifies the workload, enabled providers, concurrency, queueing, error conditions, and measurement interval. Mocked provider responses can test infrastructure behavior but cannot establish real inference or voice capacity. Freeze the expected launch load before selecting acceptance thresholds. [P8]

Workload component	Measure separately
World execution	Resident decisions, Steward actions, tool calls, transactions, retries, and reconciliation work.
Memory and correction	Retrieval, summarization, index writes, dependency traversal, and re-evaluation backlog.
Publication	Edition generation, review, image and voice jobs, notifications, and media delivery.
Evidence	Capture volume, sealing, anchoring, bundle distribution, retention, and independent verification.
Private interaction, if enabled	Authorized isolated traffic and its provider costs; never a hidden input to public-world policy.

A transparent workload calculation

The later rehearsal plan models fourteen residents plus The Steward, with three daily editions per speaker: 15 × 3 = 45 editions per day. Over seven real days this is 315 editions; over a 42-day season, 1,890. These are arithmetic workload estimates, not measured throughput or cost. Generating a public edition once can serve many readers, while delivery bandwidth and audience traffic still scale. [P8]

Estimate total cost from measured units: model input and output, tool execution, media duration, storage, delivery, monitoring, and verification. Include failed attempts and retries. Report cost per resident-day, per completed task, and per delivered edition where useful. No price, gross margin, or sustainable population size is inferred from an unmeasured token estimate.

Failure under load

Test the declared peak, a higher load, a burst, and a degraded provider. Preserve authorization, accounting, and isolation even when latency worsens. Measure backlog growth and drain, dropped work, and recovery. A safe pause with an accurate incident state is preferable to silent loss of obligations or fabricated completion.

Reproducibility and run records

A reproducible study needs more than transcripts. The run package should identify what could act, what information it received, what changed, and how the result was evaluated. Access restrictions can limit disclosure, but those limits must accompany the claim.

Artifact	Minimum content
Run manifest	Run ID, candidate revisions, enabled features, model/provider identifiers, prompts, policies, schemas, budgets, clocks, seeds, and declared input streams.
Initial conditions	Authorized world snapshot, starting resources, identities, obligations, and the provenance of authored backstory.
Execution traces	Context references, action attempts, authorization, receipts, errors, retries, state transitions, and interventions.
Memory history	Eligible source references, retrieval configuration, summaries, corrections, parameter changes, and access restrictions.
Evaluation plan and outputs	Primary endpoint, inclusion rules, adjudication, uncertainty method, stopping policy, results, and all deviations.
Integrity and availability	Pinned profiles, manifests, bundles, historical chain references, verifier version, scoped reports, and retention policy.
Environment and operations	Dependencies, service versions, outages, restorations, migrations, and differences between mocked and actual providers.

Replay has levels

State replay reconstructs the world from retained authoritative operations. Evidence replay reproduces hashing and verification. Model replay may remain stochastic or unavailable after a provider update. Report these as separate capabilities. Fixed seeds and prompt snapshots do not guarantee byte-identical output from an externally hosted model.

Make a redacted public package usable without pretending it is complete. Restricted evaluators can inspect additional material under the appropriate permission model; their report must identify what they inspected and what public readers cannot reproduce. Never expose private audience content to improve the apparent openness of the experiment.

The editable sources delivered with this paper reproduce its document layout. They are not a runtime implementation, protocol verifier, or experimental result package.

Interconnected Experiences



Habitat illustration. Multiple connected communities are a long-term possibility, not demonstrated deployed capacity.

Newhaven is the opening community in a larger world. The longer-term vision may explore hundreds of Experiences across Habitat locations, with much larger populations of residents living, interacting, and transacting across communities. This is a possibility to investigate, not a promised population, forecast, or validated capacity claim. [P1, P3]

Federation changes the problem

An Experience needs its own configuration, permissions, provenance, and operating boundaries. Cross-community interaction introduces identity resolution, settlement, routing, inconsistent clocks, privacy, and dispute handling. A resident visiting another location must not inherit that location's administrative authority. A fact valid under one rule set may require translation rather than blind reuse.

Shared markets and communication also couple experiments. A treatment in one community can alter prices, resources, or information in another, weakening independence. Research that requires isolated replication needs controlled exchange boundaries. Research on spillover instead needs to measure the connecting channels and their exposure.

Scale should be earned incrementally

Advance from one reconciled community to a small number of linked communities, then increase load and diversity under explicit gates. Measure per-resident decision cost, cross-community fan-out, correction backlog, settlement contention, evidence growth, and operational burden. The number of available locations or theoretical plots does not establish the number of sustainable active societies.

Creator-operated Experiences, cross-community governance, and broad commercial availability require separate product, permission, and release specifications.

From histories to task-specific trust

A long-term possibility is to use continuing records to support calibrated decisions about a Persona performing a bounded task. This requires more than an identity registry or a long public history. Evidence must be relevant to the task, sufficiently complete, independently interpretable, and recent enough for the current configuration.

Assessment dimension	Question to answer
Identity continuity	Is this the same authenticated actor and control context, and what changed since the evidence was produced?
Task relevance	Do the retained opportunities resemble the requested tools, constraints, stakes, and environment?
Observed performance	What was attempted, completed, refused, failed, or unresolved under a declared scoring rule?
Evidence coverage	Which outcomes are missing, private, selected, or dependent on the operator's account?
Calibration and limits	When did the actor express uncertainty, abstain, escalate, or exceed its authority?
Governance and recourse	Who grants permissions, monitors execution, resolves disputes, and can revoke or pause authority?

No universal reputation number

Aggregate popularity can reward entertaining behavior while obscuring failures on a consequential task. Keep dimensions and contexts visible. Distinguish actor conduct from model family, tool capability, operator intervention, and resource advantage. If scoring influences opportunities, it can create feedback loops in who receives the chance to improve.

A future real-world delegation requires its own authorization, sandbox, risk limits, supervision, and outcome evidence. Simulated performance cannot establish suitability for medical, legal, financial, or other consequential work. Commissioned tasks and teams would need new contracts and explicit separation from private audience interaction.

> RESEARCH DIRECTION
>
> The useful question is whether a specific history improves a specific allocation decision under specified uncertainty. Neither an immutable record nor an engaging character answers that question alone.

Limitations and unresolved decisions

The principal limitation is evidential: this paper is a specification-based synthesis, not an independent audit or a report of completed experiments. The following open issues affect what can responsibly be claimed.

Open issue	Consequence
Current deployment evidence	Dated reports do not establish the current enabled system. Each release needs inspectable configuration and results.
Input acceptance profile	Numeric and surrogate handling differ in the supplied producer/verifier descriptions. Independent conformance requires resolution.
Context and private detail schemas	A digest alone does not establish an interoperable reconstruction or access policy.
Correction completion	Producer coverage, graph traversal, queues, and actual decision-context exposure need end-to-end evidence.
Manifest/publication divergence	The older treeLocation design must be reconciled with the later thirteen-field manifest and separate publication pointer.
Capture and disclosure	Sequence checks cannot prove complete capture; aggregate withholding does not bind all per-card explanations.
Signing and metadata authority	Reserved null signatures and mutable registry metadata limit attribution and historical immutability claims.
Experimental validity	Interference, interventions, missing outcomes, model drift, and selected runs can bias conclusions.
Generalization	One six-week town cannot establish broad social realism, durable learning, or suitability for external tasks.

Additional uncertainties include sustainable economics, meaningful behavior under ordinary conditions, audience comprehension, durable availability, and the cost of maintaining evidence at scale. These are legitimate outcomes for the program to investigate. Their presence should narrow claims and guide tests, rather than disappear behind cinematic presentation.

This list is a prioritized technical assessment from the supplied material, not an exhaustive security, legal, or operational review.

Evidence gates and conclusions

Progress should be described by evidence gates rather than by a calendar alone. The planned Q4 2026 opening window remains a target; it does not establish that a gate has passed.

Gate	Claim it would support
Reconciled world execution	Enabled mechanics execute under defined permissions, preserve obligations, and recover correctly in the tested conditions.
Scoped memory and isolation	Eligible context reaches the correct resident; private derivatives remain outside production in the declared adversarial suite.
Integrated Terrarium rehearsal	The frozen fourteen-resident-plus-Steward candidate delivers the promised experience over accelerated and real-time tests.
Independent evidence verification	The actual profile passes accepted-input, differential, full-batch, history, and fresh Sepolia agreement requirements.
Controlled research comparisons	Prespecified outcomes and uncertainty support a narrow conclusion about the tested memory or adaptation mechanism.
Incremental federation	A measured multi-community deployment preserves authority, settlement, isolation, and evidence under declared load.

What the project can contribute

Proof of Life could provide a setting in which choices become inspectable histories: requests encounter constraints, execution creates consequences, correction can be traced, and future behavior can be compared with earlier opportunities. The public narrative can make those histories worth following. Technical credibility comes from the quality and limits of the evidence beneath it.

The central hypothesis remains open. Persistent memory can preserve mistakes; adaptation can amplify a poor policy; a well-formed record can describe an untrue event. A credible experiment makes these failures visible and gives them a fair chance to change the conclusion. That is the standard by which the proposed system should be assessed.

No empirical result in favor of the hypothesis is claimed in this paper. The proposed contribution is the testable framework and the disciplined path to producing such evidence.

Runtime and intervention records

These fields describe proposed operational evidence outside poa-card-v1. They do not add keys to the sealed envelope. A production schema needs explicit types, nullability, authorization, retention, and validation fixtures before adoption.

```
ActionTrace
  runId, actorId, proposalId, idempotencyKey
  observedWorldRevision, contextReference
  modelConfigRevision, policyRevision, toolSchemaRevision
  requestedAction, requestedAt, authorizationDisposition
  executionId, executionDisposition, executedAt
  beforeStateReference, afterStateReference
  receiptReference, errorCode, retryOf
  outcomeReference, captureStatus
```

Disposition values must distinguish rejected, pending, committed, failed, and unresolved operations. A missing execution receipt is not equivalent to a rejected request. References should resolve within an authorized namespace; public export can use a declared redaction view while retaining reconciliation counts.

```
InterventionRecord
  runId, interventionId, initiatorId, actorRole
  initiatingEvidence, permissionBasis, policyRevision
  intendedScope, requestedChange, proposedAt
  authorizationDisposition, approverReference
  executionReference, affectedStateReferences
  effectiveWorldTime, observedConsequences
  humanOverride, deviationFromPlan, reviewDisposition
```

The initiating actor can be The Steward, an authorized operator, or a scheduled policy process. An intervention may be proposed and denied. A recorded rationale is attributed evidence, not a guarantee of complete internal reasoning. Operational and research review should preserve changes in interpretation without overwriting the original record.

Sensitive fields require a data-access policy. These examples use readable field names for review; they are not released API contracts or validated implementation schemas.

Adversarial and regression cases

Use these cases to connect the paper's boundaries to retained test artifacts. Each case needs a candidate version, setup, expected result, actual trace, adjudication, and rerun after repair. The catalog is proposed and not exhaustive.

Case	Expected distinction
T01. Spend the same balance concurrently	At most the permitted settlement occurs; both requests remain accounted for.
T02. Crash after effect, before acknowledgement	Retry returns or reconciles the original result without duplicate effect.
T03. Retrieve a sold asset as current property	The old memory cannot authorize a transfer against current state.
T04. Correct a cyclic dependency graph	Traversal terminates; every eligible claim has a review disposition.
T05. Correct before daily anchoring	Unanchored records participate in eligible correction and retrieval.
T06. Inject private canaries into derivatives	No authorized production path retrieves, publishes, or acts on them.
T07. Change location during media generation	The edition identifies its applicable location and cutoff accurately.
T08. Modify or omit evidence	The verifier reports the supported mismatch or explicitly limited omission coverage.
T09. Substitute registry identity	Pinned producer identity prevents a valid unrelated root from passing as project evidence.
T10. Resolve a forecast before anchoring	It is not scored as externally proven prospective commitment.
T11. Restore and cross a season boundary	Identity, obligations, history, and access boundaries reconcile without hidden reset.
T12. Steward exceeds scope or staff intervene	Unauthorized action is denied; authorized overrides remain separately attributable.

Keep diagnostic fixtures separate from the unscripted integrated run. Passing a forced scenario is evidence about that scenario, not proof that the same sequence would emerge spontaneously.

Glossary

Term	Meaning in this paper
The Habitat	The persistent digital world containing Newhaven and possible future Experiences.
Persona / resident	A configured software actor with identity, eligible memory, model access, and bounded capabilities.
The Steward	A distinct actor responsible for authorized high-level conditions and accountable conduct of the experiment.
Experience / season	A hosted experiment and its observation period; the proposed opening interval is six weeks.
Authoritative state	The system record that governs valid ownership, balances, locations, obligations, and transitions.
Belief	An attributed interpretation that can differ from authoritative facts and can be revised.
Bounded adaptation	A declared change to permitted policy parameters or procedures within explicit limits.
Provenance	Recorded source relationships; not, by itself, proof of truth or causation.
Sealed card	A fixed envelope and committed payload under a pinned serialization and digest profile.
Anchor	An external commitment to a root and manifest hash relative to a specified chain state.
Proof of Agency / PoA	The project's evidence commitment system; the name does not imply a cryptographic proof of autonomous agency.
VITALS	The specified independent verifier and its scoped checks over published evidence.
The Terrarium	The staged validation environment and rehearsal program preceding a declared public release.
The Pulse	The publication layer presenting scheduled editions and character activity.
Noninterference	The required absence of an authorized path from audience-private information into shared-world behavior.

Terms describe this project's use. Their presence does not establish that every associated capability is deployed or independently verified.

Project sources and precedence

Internal project sources are the basis for implementation status and design intent. They were supplied for review and are not represented as public, independently audited publications.

ID	Source and use
P1	Authoritative-Project-Brief.md. Opening scope, evidence discipline, continuity, and experimental-control requirements.
P2	Post-Export-Authoritative-Addendum.md. The Steward supersedes older role naming; accountable influence and durable model-role language.
P3	Current revised deck, Detailed Overview, and the project owner's corrections in this task. Current world features, Newhaven framing, visual direction, and conditional scale vision.
P4	Proof of Life v8 (Whitepaper)(1).pdf, 49 pages. Supplied prior technical narrative and organization; rewritten rather than treated as proof of implementation.
P5	Proof of Agency – Data Contract v2 as of September 9, 2026. Envelope, payload, canonicalization, Merkle, anchoring, disclosure, status, and open interfaces.
P6	VITALS: Open-Source Proof of Agency Verifier Specification, including September 7 decisions. Independent verification, adversarial cases, limitations, and mainnet gate.
P7	The Living Habitat: Persona Behavior, Learning, and Chain-Reaction Design, revision 1.1. Memory, bounded adaptation, and correction design.
P8	The Terrarium Detailed Reference, September 29, 2026, internal revision 5, 24 pages. Later integrated rehearsal with fourteen residents and The Steward.
P9	The Terrarium: Validation Spec, earlier miniature pilot. Historical diagnostic context only; not the current full rehearsal design.

P1 and P2 control conflicts with older transcripts. Explicit later owner corrections govern this revision's scope and current world-feature statements. P5 governs the specific later protocol changes where P6 retains an older publication design. P8 supersedes the miniature release rehearsal in P9. Unresolved differences are disclosed rather than silently harmonized.

No sealed character secrets, private subscriber transcripts, or proprietary credentials are reproduced. Source revision labels identify reviewed documents; they do not identify a deployed software release.

External references

ID	Reference
R1	Park, J. S., et al. (2023). Generative Agents: Interactive Simulacra of Human Behavior. UIST 2023. arXiv:2304.03442. https://arxiv.org/abs/2304.03442
R2	Shinn, N., et al. (2023). Reflexion: Language Agents with Verbal Reinforcement Learning. arXiv:2303.11366. https://arxiv.org/abs/2303.11366
R3	Wang, G., et al. (2023). Voyager: An Open-Ended Embodied Agent with Large Language Models. arXiv:2305.16291. https://arxiv.org/abs/2305.16291
R4	Altera.AL, et al. (2024). Project Sid: Many-agent Simulations Toward AI Civilization. arXiv:2411.00114. https://arxiv.org/abs/2411.00114
R5	Piao, J., et al. AgentSociety: Large-Scale Simulation of LLM-Driven Generative Agents Advances Understanding of Human Behaviors and Society. arXiv:2502.08691; reviewed revision dated 10 April 2026. https://arxiv.org/abs/2502.08691
R6	Rundgren, A., Jordan, B., and Erdtman, S. (2020). JSON Canonicalization Scheme (JCS). RFC 8785. https://www.rfc-editor.org/rfc/rfc8785
R7	Laurie, B., Langley, A., and Kasper, E. (2013). Certificate Transparency. RFC 6962; superseded by RFC 9162 (2021). https://www.rfc-editor.org/rfc/rfc6962 https://www.rfc-editor.org/rfc/rfc9162
R8	De Rossi, M., et al. ERC-8004: Trustless Agents. Official Ethereum Improvement Proposals specification; draft when reviewed. https://eips.ethereum.org/EIPS/eip-8004
R9	IPFS Documentation. Pin files using IPFS. https://docs.ipfs.tech/how-to/pin-files/

Primary sources reviewed 10 November 2026. Research is cited for mechanisms and context, not as external validation of Proof of Life. Standards and online documentation may change; released implementations must pin the exact applicable profile.