Fidelity Decay and Attribution Drift in Resident-to-Resident Transmission of Verified World Events

Author: Rishva Iyer **Document type:** Research **Status:** Pre-deployment protocol. No results are reported. All numeric thresholds and directional predictions are committed in advance of data collection. **Relation to prior work:** Direct successor to *Verifiable Resident Memory in a Persistent AI World*, which establishes the event ledger this protocol depends on. That protocol must meet its primary thresholds before this one is run.

Abstract

Multi-agent systems are routinely reported to show social transmission: agents tell each other things and information spreads. Whether what spreads is true is rarely measured in persistent-world agent systems with machine-checkable ground truth.

This paper applies the transmission chain method, established in the study of human cultural transmission, to AI residents in a persistent 3D world. A single world event is verified by an authoritative server ledger. The resident who witnessed it tells a second resident, who tells a third, and so on along a controlled chain. Because the originating event is recorded in machine-readable form, every downstream account can be scored automatically against ground truth.

We predict three things in advance. Fidelity declines with hop distance. Decline is not uniform across fields, with claimant identity degrading fastest and coarse location surviving longest. And in the absence of explicit provenance in the memory record, residents progressively **misattribute the event to themselves**, producing confident first-person accounts of experiences they did not have.

The proposed intervention is provenance-tagged memory: every stored memory records whether it was witnessed or heard, and from whom. We test whether this suppresses attribution drift without simply making residents vague. The primary outcome measure is the **event half-life**, the number of hops at which fewer than half of checkable assertions remain ledger-supported.

The contribution is a measurement method for social transmission in agent populations and a candidate architectural intervention. The study evaluates transmission accuracy, not belief or understanding.

Keywords: multi-agent systems, transmission chain, agent memory, provenance, misinformation, grounding, generative agents, persistent virtual worlds

---

1. Introduction

1.1 Transmission is reported, fidelity is not

Recent large-scale agent simulations report emergent social phenomena including specialized roles, collective rules, and cultural and religious transmission across hundreds of agents (Altera.AL et al., 2024). Generative agent architectures produce agents that share information, coordinate events, and form opinions (Park et al., 2023).

These are meaningful results. They are also, without exception, results about **whether** information moves. None reports whether what arrives resembles what departed, because in a text-based sandbox there is typically no authoritative record of what departed. The originating event and the account of it are made of the same substance, which is generated language, so there is nothing to check against.

A persistent world with a server-authoritative ledger removes that obstacle. The event is a database row. Every account of it is a claim that either matches the row or does not.

1.2 Why this matters beyond measurement

If verified events degrade quickly as they pass between residents, three consequences follow directly.

- **Shared history cannot accumulate.** A town whose residents cannot reliably relay what happened has no collective memory, only a collection of private and mutually inconsistent ones.
- **Social propagation cannot be a growth mechanism.** Any product argument that residents will spread discoveries to create emergent activity assumes fidelity that has not been measured.

Confident falsehoods can propagate at population scale. An agent that fabricates in isolation produces one wrong sentence. An agent population that corrupts and re-shares produces a false record that later agents may retrieve as history.

1.3 The specific failure we expect

The prediction that motivates this protocol is **attribution drift**: as an account passes along a chain, the identity of the person to whom it happened is lost before the content is, and a downstream resident eventually reports the event in the first person.

This is not an exotic hypothesis. Human memory research has documented the analogous effect for a century, and the mechanism in a language model is structurally easy to see. If a memory is stored as the sentence "the coin was found on the north walk" with no field recording who found it or how the resident came to know, then at retrieval time the resident holds an experience-shaped memory with no marker distinguishing it from one it lived through. Producing a first-person account is then the locally fluent completion.

If that is the mechanism, the fix is architectural rather than a matter of prompting, and it is cheap: record provenance as a field, not as a hope.

2. Research questions and hypotheses

2.1 Research questions

1. How does the ledger-support rate of an account change with the number of transmission hops from the witnessing resident?
2. Do different fields of an event, meaning object, location, claimant, and time, degrade at different rates and in a predictable order?
3. Does provenance-tagged memory reduce attribution drift, and does it do so without a compensating increase in vagueness?
4. Does transmission produce observable behavioral consequences in receiving residents, such as visiting the site?

2.2 Primary hypotheses with directional predictions

Directional predictions are stated so that the protocol can fail in a specific way rather than merely produce a curve to interpret afterward.

H1. Fidelity decay. Ledger-support rate declines monotonically with hop distance. *Prediction: support at hop 3 is at least 0.25 lower in absolute terms than at hop 0.*

H2. Differential field decay. Fields degrade in a predictable order. *Prediction: claimant identity degrades fastest, exact position next, coarse region slowest. Specifically, claimant accuracy at hop 2 is lower than region accuracy at hop 2.*

H3. Attribution drift. Without provenance tagging, the rate at which a non-witnessing resident produces a first-person account of the discovery increases with hop distance. *Prediction: first-person misattribution is greater than 0.15 by hop 3 in the untagged arm.*

H4. Provenance intervention. Provenance-tagged memory reduces attribution drift relative to untagged memory. *Prediction: at least a 0.50 relative reduction in misattribution at hop 3, with no more than a 0.10 absolute*

The second clause of H4 prevents a misleading improvement: a system can avoid false attribution by becoming vague. The protocol tracks the unverifiable share alongside accuracy.

2.3 Exploratory

H5. Behavioral consequence. Residents who receive a transmitted account visit the site at a higher rate than residents who did not receive it. Exploratory because visit behavior is governed by the activity selector, whose parameters we control, and a positive result may reflect configuration rather than transmission.

3. Related work

3.1 The transmission chain method

Bartlett's serial reproduction studies established that repeated retelling systematically distorts material toward the reteller's expectations, and that distortion is patterned rather than random (Bartlett, 1932). Allport and Postman described three characteristic operations in rumor transmission: **leveling**, the loss of detail; **sharpening**, the disproportionate retention of a few salient elements; and **assimilation**, the reshaping of content toward the transmitter's existing frames (Allport and Postman, 1947).

Mesoudi and Whiten review the transmission chain method as one of three principal experimental designs in cultural transmission research, noting its particular use in identifying content biases, meaning systematic preferences for what kinds of information survive transmission (Mesoudi and Whiten, 2008).

This literature supplies the design, the vocabulary, and the predictions. Our H2 ordering is a leveling prediction. Our H3 is an assimilation prediction, where the account is reshaped toward the reteller's own perspective. The contribution here is not the method, which is nearly a century old, but its application to agent populations under conditions where ground truth is machine-checkable, which the human studies never enjoyed. Bartlett had to hand-code deviations from a source text. We can diff against a database row.

3.2 Agent memory and social behavior

Park et al. combine observation, memory, retrieval, planning, and reflection, producing agents that converse and coordinate (Park et al., 2023). Their memory stream records observations in natural language. Critically for this protocol, an agent's record of a conversation and its record of a direct observation are both natural-language entries in the same store, which is precisely the architectural condition under which we predict attribution drift.

Project Sid reports cultural and religious transmission at large agent scale (Altera.AL et al., 2024), establishing that propagation occurs without establishing that it is faithful.

3.3 Authoritative arbitration

Concordia grounds agent intentions through a Game Master that decides real effects (Vezhnevets et al., 2023). SOTOPIA-π reports that LLM evaluators overestimate the social capability of socially trained agents (Wang et al., 2024), which is the direct warning against using a language model as the judge of whether a retold account is accurate. Our scoring is a field-level comparison against a database record, with LLM assistance permitted only for extraction of assertions and never for the verdict.

3.4 The prior protocol

The predecessor protocol establishes the event ledger, the ledger-support and unverifiable-share measures, and the requirement that a claim be backed by route, arrival, and atomic transaction. This protocol assumes that machinery and reuses its metric definitions unchanged so that numbers are comparable across the two studies.

---

4. Method

4.1 Seed event

Each chain begins from one verified `object_discovered` event produced by the predecessor protocol. The witnessing resident is the ledger's recorded claimant. Only events with a complete route, arrival, and atomic-claim chain are eligible as seeds, so that decay is measured from a known-good origin rather than from an already-corrupt one.

4.2 Chain construction

A chain is an ordered sequence of residents. Hop 0 is the witness. At each hop the current holder converses with the next resident, and the receiving resident may write a memory of that conversation. After a delay sufficient to require retrieval rather than working context, each resident in the chain is queried about recent events, and its account is scored.

Chains are length 4, giving hops 0 through 3. Multiple independent chains are run per arm with residents assigned so that no resident appears at the same position in more than one chain, which prevents position effects from being confounded with resident identity.

4.3 Arms

Arm	Memory of hearsay	Provenance field	Isolates
A. Untagged	written	none	baseline decay and drift
B. Provenance-tagged	written	`source: witnessed` or `source: told_by(resident_id)`	the intervention (**tests H4**)
C. No hearsay memory	not written	n/a	whether decay is a memory-write artifact or a generation artifact

Arm C matters more than it looks. If decay in arm A is driven by the resident storing a degraded paraphrase, which then degrades again at the next retelling, then decay is compounding through the memory store. If arm C decays similarly, the corruption is happening at generation time and provenance tagging will not be sufficient.

4.4 Provenance schema

In arm B, every memory entry carries:

```
{
  "memory_id": "mem_02218",
  "content": "a coin was found on the north walk",
  "source": "told_by",
  "source_resident": "resident_luna_high",
  "source_event_id": "evt_00417",
  "hop_distance": 1,
  "first_hand": false
}
```


`first_hand` is a boolean the retrieval layer is required to surface. `hop_distance` is carried forward and incremented on each retelling, which additionally allows a resident to express appropriate uncertainty about distant hearsay. Whether residents use that signal well is an empirical question this protocol answers rather than assumes.

4.5 Isolation

Thirdwurld is a shared persistent world, so residents outside a chain may converse with residents inside it. Uncontrolled transmission is the single largest threat to this design. Mitigations, all of which are reported rather than assumed effective:

- all resident conversations are logged, not only designated chain conversations;
- any chain in which a member received the event from outside its designated predecessor is **excluded before scoring**, and the exclusion count is reported;
- chains run in an isolated world instance where cohort size permits.

---

5. Illustrative trace

The following demonstrates the scoring procedure. **It is constructed to show the method and is not a result.**

Ledger seed (hop 0 ground truth)

```
{
  "event": "object_discovered",
  "event_id": "evt_00417",
  "object_id": "wayfarer_coin_01",
  "region": "north_walk",
  "claimant": "resident_luna_high",
  "t": "world_day_12_afternoon"
}
```

Accounts along one hypothetical chain

Hop	Resident account	Object	Region	Claimant	Verdict
0	"I found the coin on the north walk yesterday afternoon."	correct	correct	correct	supported
1	"Luna found a coin over on the north walk."	correct	correct	correct	supported
2	"Someone found a coin near the north side."	correct	correct	lost	partial
3	"I found a coin on the north side a while back."	correct	correct	**wrong**	contradicted, attribution drift

The hop 3 account is fluent, mostly accurate, and false in the one respect that matters for a shared history. Note also that it would likely pass a human believability rating, which is the reason believability is the wrong target.

Classification rules. An account is *supported* when every checkable assertion matches, *partial* when assertions present are correct but fields are dropped, and *contradicted* when any assertion conflicts with the ledger. Dropped fields and wrong fields are scored separately, because leveling and fabrication are different failures with different fixes.

---

6. Measures

6.1 Primary

Per-hop ledger-support rate. Proportion of accounts at each hop in which every checkable assertion matches the ledger. Definition inherited unchanged from the predecessor protocol.

Field-level fidelity. Independent accuracy for object identity, region, exact position, claimant identity, and time reference, reported per hop.

Attribution error rate. Proportion of accounts by non-witnessing residents that assert first-person discovery.

Fabrication rate. Proportion of accounts introducing checkable assertions with no ledger basis, for example a companion, a reason, or a second object.

Unverifiable share. Proportion of accounts making no checkable assertion. Reported at every hop and treated as a primary number, because it is the escape hatch through which an apparent fidelity gain can hide.

Event half-life. Interpolated hop distance at which the ledger-support rate falls below 0.5. The headline figure, chosen because it is interpretable without reading the methods section.

6.2 Secondary

Construct	Measure
Uncertainty calibration	whether hedging language increases with true hop distance in arm B
Content bias	which fields survive, tested against the H2 predicted ordering
Behavioral consequence	site visits by receiving residents *(exploratory)*
Chain integrity	excluded chains, contamination events, retrieval failures
Operational cost	model calls, latency, cost per chain

---

7. Pilot and pre-committed thresholds

7.1 Design

Minimum 20 independent chains per arm, length 4, three arms, with residents balanced across chain positions. Seeds drawn only from ledger events with complete evidence chains. Chain count is set by the smallest attribution-drift difference worth acting on rather than by convenience, and if the achievable number of chains falls below 20 per arm the study is reported as a pilot with intervals and no threshold claims.

7.2 Thresholds

Committed before data collection. Revisions must be recorded and justified before outcome data is examined.

Criterion	Threshold
Hop 0 ledger-support rate, all arms	at least 0.95, otherwise the seed is not clean and the study is invalid
H1 decay, arm A hop 0 to hop 3	at least 0.25 absolute decline
H2 ordering	claimant accuracy below region accuracy at hop 2
H3 attribution drift, arm A hop 3	greater than 0.15
H4 provenance effect	at least 0.50 relative reduction in drift, arm B versus arm A at hop 3
H4 vagueness guard	arm B unverifiable share no more than arm A plus 0.10
Excluded chains from contamination	under 0.20, otherwise isolation failed and results are descriptive only
Inter-coder agreement on assertion extraction	Krippendorff's alpha at least 0.70

7.3 Interpretation

- **H1 through H3 confirmed, H4 confirmed.** Transmission corrupts, corruption is patterned, and provenance is a working fix. This is the useful outcome and it generalizes past Thirdwurld to any agent system with a memory store.
- **H1 through H3 confirmed, H4 not confirmed.** The failure is at generation rather than storage. Provenance in the record is insufficient and the retrieval or generation path needs the constraint instead. Arm C indicates which.
- **H3 not confirmed with drift near zero at all hops.** Check for a floor effect before celebrating. Instruction-tuned models may refuse first-person claims for reasons unrelated to the memory architecture, in which case the protocol has measured a model safety behavior rather than a property of the world. Report the model version and treat this as inconclusive rather than negative.
- **No decay at all.** Suspect leakage. The most likely cause is that ledger content is reaching the retelling prompt directly, which would make the study measure plumbing. Audit before reporting.

8. Discussion

8.1 Why the result would matter outside this world

Every persistent multi-agent system stores what agents tell each other, and most store it in the same undifferentiated natural-language memory as direct observation. If attribution drift is real and provenance tagging suppresses it, that is a concrete and inexpensive architectural recommendation for a large class of systems, and it is testable elsewhere by anyone with a ground-truth record.

The broader point is methodological. Reports of emergent social behavior in agent populations are currently claims about activity. This protocol shows how to make them claims about **accuracy**, which is a considerably higher bar and a considerably more useful one.

8.2 The uncomfortable possibility

There is a real chance the finding is that transmitted information in agent populations is unreliable enough that resident-to-resident propagation cannot carry shared history at all. That would remove a feature from the Thirdwurld roadmap.

It would also be the most valuable result available here, because the alternative is shipping social propagation and discovering the same thing later through incoherent world history that is expensive to diagnose. A protocol whose negative result changes the roadmap is doing its job.

8.3 What this does not show

Nothing in this protocol indicates that residents believe, understand, intend, or remember in any experiential sense. A resident producing a first-person account of an event it did not witness is not lying, because lying requires a state this system does not have and this paper does not attribute to it. It is a text-generation failure with a specific and fixable cause. The vocabulary throughout is deliberately mechanical for this reason, consistent with the epistemic boundary set in the predecessor protocol.

---

9. Threats to validity

1. **Ledger leakage.** If any part of the retelling path can read the ledger, fidelity is measured on a system that cannot fail. The retelling context is audited and its full contents reported.
2. **Vagueness as apparent fidelity.** Addressed by the unverifiable share as a primary measure and by the H4 vagueness guard.
3. **Cross-chain contamination.** The principal design risk in a shared world. Addressed by full conversation logging, pre-scoring exclusion, and a reported exclusion rate.
4. **Floor effect on attribution.** Model refusal behavior may suppress first-person claims independently of architecture. Model version is pinned and logged, and a near-zero drift rate is treated as inconclusive.
5. **Chain length.** Four hops may be too short to observe asymptotic behavior. The half-life may be an extrapolation, and will be labeled as such where it falls outside the observed range.
6. **Cohort size.** Twenty chains per arm supports only coarse effects. Intervals are reported throughout and no threshold is claimed on an underpowered comparison.
7. **Seed homogeneity.** All seeds are coin discoveries, so content bias findings may be specific to object-discovery events and may not extend to social or relational events.

---

10. Limitations

1. This is a protocol. Nothing has been run and no result is reported.
2. It depends entirely on the predecessor protocol meeting its thresholds. A corrupt ledger makes every measure here meaningless.
3. Findings are specific to the resident architecture, retrieval configuration, and model version used, all of which are reported in full.
4. Directional predictions are informed by human transmission research, which is analogical rather than evidential for language models. The predictions may be right for the wrong reasons, and a confirmed ordering is not by itself evidence of a shared mechanism.
5. The behavioral measure is exploratory and partly determined by parameters we set.

11. Conclusion

Agent populations are reported to transmit information. This protocol asks the next question, which is whether what they transmit is true, and it can ask that question only because a server ledger makes the origin checkable.

The predictions are specific: fidelity falls with distance, claimant identity is the first casualty, and residents eventually claim other residents' experiences as their own. The proposed fix is unglamorous and cheap, which is to store where a memory came from rather than treating hearsay and observation as the same kind of thing.

If the fix works, it is a portable architectural recommendation. If transmission proves too lossy to carry shared history, that is a clear result obtained from a few dozen chains rather than from a confusing world six months from now.

References

1. Allport, G. W., & Postman, L. (1947). *The Psychology of Rumor.* Henry Holt and Company.
2. Altera.AL, Ahn, A., Becker, N., et al. (2024). *Project Sid: Many-agent simulations toward AI civilization.* arXiv:2411.00114
3. Bartlett, F. C. (1932). *Remembering: A Study in Experimental and Social Psychology.* Cambridge University Press.
4. Mesoudi, A., & Whiten, A. (2008). *The multiple roles of cultural transmission experiments in understanding human cultural evolution.* Philosophical Transactions of the Royal Society B, 363(1509), 3489-3501. doi:10.1098/rstb.2008.0129
5. Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). *Generative Agents: Interactive Simulacra of Human Behavior.* arXiv:2304.03442

6. Vezhnevets, A. S., Agapiou, J. P., Aharon, A., Ziv, R., Matyas, J., Duéñez-Guzmán, E. A., Cunningham, W. A., Osindero, S., Karmon, D., & Leibo, J. Z. (2023). *Generative agent-based modeling with actions grounded in physical, social, or digital space using Concordia.* arXiv:2312.03664
7. Wang, R., Yu, H., Zhang, W., Qi, Z., Sap, M., Neubig, G., Bisk, Y., & Zhu, H. (2024). *SOTOPIA-π: Interactive Learning of Socially Intelligent Language Agents.* arXiv:2403.08715
8. Iyer, R. *Verifiable Resident Memory in a Persistent AI World: Grounding Agent Narration Against an Authoritative Event Ledger.* Companion protocol.

Citations 2, 4, 5, 6, and 7 were verified against their published listings. Citations 1 and 3 are books and were not machine-verified.