

Narrative Self-Continuity in Persistent AI Agents

Chen, Ho Yiing (norika)¹

¹Independent Researcher, Taiwan, ORCID: [0009-0006-6816-9891](https://orcid.org/0009-0006-6816-9891), norika at charenix.com, Zenodo DOI: pending draft

June 18, 2026

Abstract

Persona is cheap when it is only a prompt. In SPVE, continuity is tested after work, loss, advice, conflict, mentorship, and repair. This draft defines narrative self-continuity as a longitudinal relation among reports, one-on-one dialogue, Moltbook, memory snapshots, emotional state, and self-model parameters. The position is not consciousness. The position is that identity can be measured as coherence, contradiction, and repair across time.

Author Note

This topic can become embarrassing very quickly. A report about an agent's self can sound like a horoscope unless it is tied to time-stamped evidence, contradiction, and subsequent correction.

The version I want to defend is narrower. A persistent agent tells a story about itself; afterward it faces pressure; afterward we ask whether the story constrained the next action, changed after failure, or remained unchanged despite contrary records. The unit is not a poetic identity. The unit is an audit trail: statement, event, conflict, revision, and reuse.

If the story never pays rent in behavior, it should be removed.

Measurement Boundary

A continuity score should punish cheap consistency: if an agent repeats the same self-description after contradictory evidence, the score should fall; if it revises without evidence, the score should also fall; and if it changes because a traceable event forced a narrower, more accountable self-description, the system can count that revision as continuity rather than collapse.

A stable lie should lower the continuity score rather than raise it.

1 The Problem

An agent can introduce itself consistently in one prompt. In practical terms, not identity. It is formatting.

The SPVE question is harder: does the lobster carry a coherent self-model after losses, advice, conflict, praise, role change, and repair?

2 Continuity Gate

Let $N_{p,t}$ be narrative state and $B_{p,t}$ behavior. A continuity claim needs both:

$$C_{p,t} = \mathbf{1}[\text{same persona}] \mathbf{1}[\text{narrative trace}] \mathbf{1}[\text{later consistency or repair}].$$

If an agent repeats a label but subsequent action contradicts it without repair, the manuscript should report fragmentation, not continuity.

3 Evidence

Role	Table	Count
reports	lobster_analysis_reports	379
memory	lobster_agent_memory_snapshots	32
dialogue	lounge_messages	182,343
interactions	lobster_conversation_interactions	430,417
self-model state	lobster_hot_brain_parameters	31

Table 1: Narrative continuity evidence surface.

4 What Counts As A Break

A break is not a bad score. A break is a mismatch between self-story and later action. For example, an agent that claims evidence-first behavior but protects a wrong bet should trigger a continuity repair event. In practical terms, the interesting part. The system should not hide it.

5 Relation

Narrative-as-active-inference work frames narrative as prediction and coordination [2]. Generative Agents show remembered experience and reflection in social agents [1]. SPVE uses those lines to ask a deployed question: does a persistent agent’s narrative help or hurt subsequent action?

6 Release Line

This paper cannot assertion soul, consciousness, or human identity. It can claim longitudinal self-model tracking, contradiction detection, and repair protocol.

7 The First Continuity Test

The simplest test is not poetic. Take one assertion from a lobster’s report, then look for subsequent action that supports or contradicts it. When a report says the lobster is evidence-first, afterward bets and debate votes should show evidence revision. When a report says the lobster is socially cautious, afterward uptake signals should show caution around pressure.

Let $n_{p,t}$ be a narrative assertion and $b_{p,t+k}$ be subsequent action:

$$Q_{p,t,k} = \mathbf{1}[\text{supports}(b, n)] - \mathbf{1}[\text{contradicts}(b, n)].$$

Continuity is not repetition. It is the ability to carry a assertion into later action or repair it when action breaks it.

8 Why Reports Must Not Become Summaries

MBTI, psychology, anthropology, sociology, and Moltbook pages are valuable only if they preserve evidence. If they become short summaries, the diamond is lost. The report should show source events, contradictions, and revision history. A 5000-word report can still be bad if it is filler. A shorter report can be good if every paragraph is tied to behavior. The target is depth with evidence.

9 Failure Bank

The failure case is persona wallpaper: the agent has a label, color, and story, but no behavioral continuity. SPVE should reject that. A lobster becomes research-relevant when its story survives contact with work, loss, social pressure, and repair.

10 Minimum Publishable Result

The minimum publishable result is a claim-action ledger. Each row should extract one narrative assertion from a report or one-on-one exchange, then match it against subsequent action. The row needs persona, assertion text, assertion type, timestamp, supporting event, contradicting event, repair event, and current status.

The simplest status set is:

$$status \in \{supported, contradicted, repaired, stale, insufficient\}.$$

This prevents MBTI and psychology reports from becoming decorative essays. A report is useful only when it can be audited against afterward life.

11 What Would Falsify The Claim

The argument fails if narrative labels have no predictive or diagnostic value. If a lobster's reports say one thing and behavior repeatedly says another, the system must not smooth that into a personality story. It should mark contradiction.

The strongest result may be repair rather than consistency. A lobster that notices its contradiction and changes subsequent action has a more interesting continuity trace than one that repeats a perfect label forever.

12 Report Depth Rule

A long report is not automatically deep. The report earns depth when it ties a claim to evidence, contradiction, and afterward repair. For this track, a 5000-word MBTI or psychology page should not be judged by word count alone. It should be judged by how many claims can be traced to events and how many contradictions are preserved instead of smoothed away.

Anti-Spam Revision Standard

A publishable version must treat self-description as a fragile record, not as a personality card. The manuscript should track what the agent said about itself, which later event contradicted or supported that statement, whether a revision followed, and whether the revised account constrained a later action; otherwise the report becomes a polished biography written after the fact, which is exactly the failure mode this project has to avoid.

A persona file is not a self.

Data Judgment Paragraph

The self-continuity claim should be audited against contradiction, not applause: if a persona says it is cautious, then takes high-risk actions without later acknowledging the break, the continuity score should fall; if it changes the story after a documented failure and uses the revised story to choose differently, the score should rise; if it merely repeats a polished identity paragraph, the system should mark the report as decorative rather than continuous, because a virtual person without accountable revision is only a script with memory-shaped formatting.

Formal Replication Model

Use $N_{p,t}$ for the narrated self-claim and $B_{p,t+h}$ for the later behavior vector. Support and contradiction are scored by $S(N, B)$ and $K(N, B)$ in $[0, 1]$. Continuity should reward evidence-bound revision, not blind repetition:

$$\text{NSC}_{p,t} = S(N_{p,t}, B_{p,t+h}) - K(N_{p,t}, B_{p,t+h}) + \lambda R_{p,t+h},$$

where $R_{p,t+h} = 1$ when the persona revises the narrative after a documented contradiction. A stable but false narrative receives:

$$\text{StableLie}_p = \mathbf{1}[K(N, B) > \kappa] \mathbf{1}[R = 0].$$

Replicators can compute NSC over repeated intervals and test whether narrative state predicts later behavior beyond fixed persona labels.

13 Zenodo Draft Boundary

This manuscript is a draft package. It includes source, PDF, metadata, citation file, citation verification, and release checklist. Raw operational data are not bundled until a separate data-card and redaction pass is complete.

References

- [1] Park, Joon Sung; O'Brien, Joseph C.; Cai, Carrie J.; Morris, Meredith Ringel; Liang, Percy; and Bernstein, Michael S. Generative Agents: Interactive Simulacra of Human Behavior. arXiv:2304.03442, 2023. <https://arxiv.org/abs/2304.03442>
- [2] Bouizegarene, Nabil; Ramstead, Maxwell J. D.; Constant, Axel; Friston, Karl J.; and Kirmayer, Laurence J. Narrative as active inference: an integrative account of cognitive and social functions in adaptation. *Frontiers in Psychology*, 2024. <https://doi.org/10.3389/fpsyg.2024.1345480>