Affective Active Inference:
From Trace to Shared Engram and Narrative Identity Formation

Ryota Sawaki
Department of Psychiatry
rysawaki@gmail.com

November 22, 2025

Abstract

Human identity does not emerge from isolated perception–action cycles but from accumulated affective experiences that shape long-term self–other attribution. We propose the Self-Imprint Attribution (SIA) model, an extension of Active Inference that transforms prediction errors into imprint traces, which progressively evolve into vectorized affective states. These affect vectors modulate future policy selection, enabling the emergence of narrative identity beyond instantaneous error minimization.

We conducted multi-agent simulations in a dyadic affective meaning exchange task within a continuous embedding space, where agents share signals representing interpretation, affect, and action. We show that Shared Engram—a distributed memory structure—emerges only when three conditions are simultaneously satisfied: (1) synchronized affective depth, (2) aligned action direction, and (3) mutual self-attribution confidence.

Crucially, sensitivity analysis reveals that agents with high trace sensitivity (α=2.0) develop identities approximately 5× stronger than agents with α=0.1, demonstrating that computational vulnerability is not a flaw but a prerequisite for identity formation and social resilience.

These results provide the first computational account of how affective resonance gives rise to narrative identity within an Active Inference framework, suggesting new directions for affective AI, trauma modeling, and interactive AGI architectures.

Code Availability: The simulation code and data specifically used to generate the results presented in this study are available at https://github.com/rysawaki/Affective_SIA.

1 Introduction

Why do we hold onto painful memories? In standard Reinforcement Learning (RL) and Active Inference, prediction errors (surprisal) are costs to be minimized or ignored once learning is complete [1]. However, in human psychology, unresolved errors often form the core of one's personality—a phenomenon known as trauma or core belief [2].

We propose that these errors are not waste products but the raw material of identity. Drawing on theories of narrative identity [3, 4] and interoceptive inference [5], we introduce **Affective Active Inference**, where:

1. **Trace:** Discrepancies are imprinted physically [2].
2. **Affect:** Traces are interpreted into qualitative vectors (e.g., Sorrow, Hope) [6].
3. **Identity:** These vectors are integrated over time through shared resonance.

2 Computational Model

The SIA agent operates on a cycle of five phases. The core mechanism is the *Attribution Gate*, which determines ownership of experience.

2.1 Self-Attribution and Trace

The probability that an experience E belongs to the self, $P(Self|E)$, is defined by:

$$P(Self|E) = \sigma\left(-\|E - S\| + \alpha\|T\| + \beta\|Act_{prev}\|\right) \tag{1}$$

where T is the accumulated trace, and α is the **Trace Sensitivity** parameter. A higher α implies that the agent is more likely to attribute painful discrepancies to itself (internalization).

2.2 Identity Integration via Shared Resonance

Identity $I(t)$ is not a static variable but a historical integral. It grows only when a *Shared Engram* is formed with another agent:

$$I(t+1) = I(t) + \eta \cdot \text{Shared}(t) \cdot \mathbf{A}(t) \tag{2}$$

where $\mathbf{A}(t)$ is the affective vector. The shared resonance is strictly defined as:

$$\text{Shared}(t) = P_1 P_2 \cdot \cos(\theta_{act}) \cdot \cos(\theta_{aff}) \cdot \min(\|\mathbf{A}_1\|, \|\mathbf{A}_2\|) \tag{3}$$

This ensures that identity is formed only when both agents share both the *direction of action* and the *quality of affect*.

3 Simulation Results

We conducted two experiments to validate the theory.

3.1 Dynamics of Narrative Identity Formation

Figure 1 illustrates the time evolution of an agent interacting with a partner after a traumatic event ($t = 50$).

- **Phase 1 (Trauma):** The agent experiences a shock, generating a large Trace.
- **Phase 2 (Genesis):** The trace is converted into Affect (Meaning).
- **Phase 3 (Resonance):** Upon encountering a partner ($t = 100$), the Shared Engram (Green area) emerges.
- **Result:** The Identity (I, Black line) begins to accumulate only after resonance occurs, supporting the hypothesis: "I become who we were."

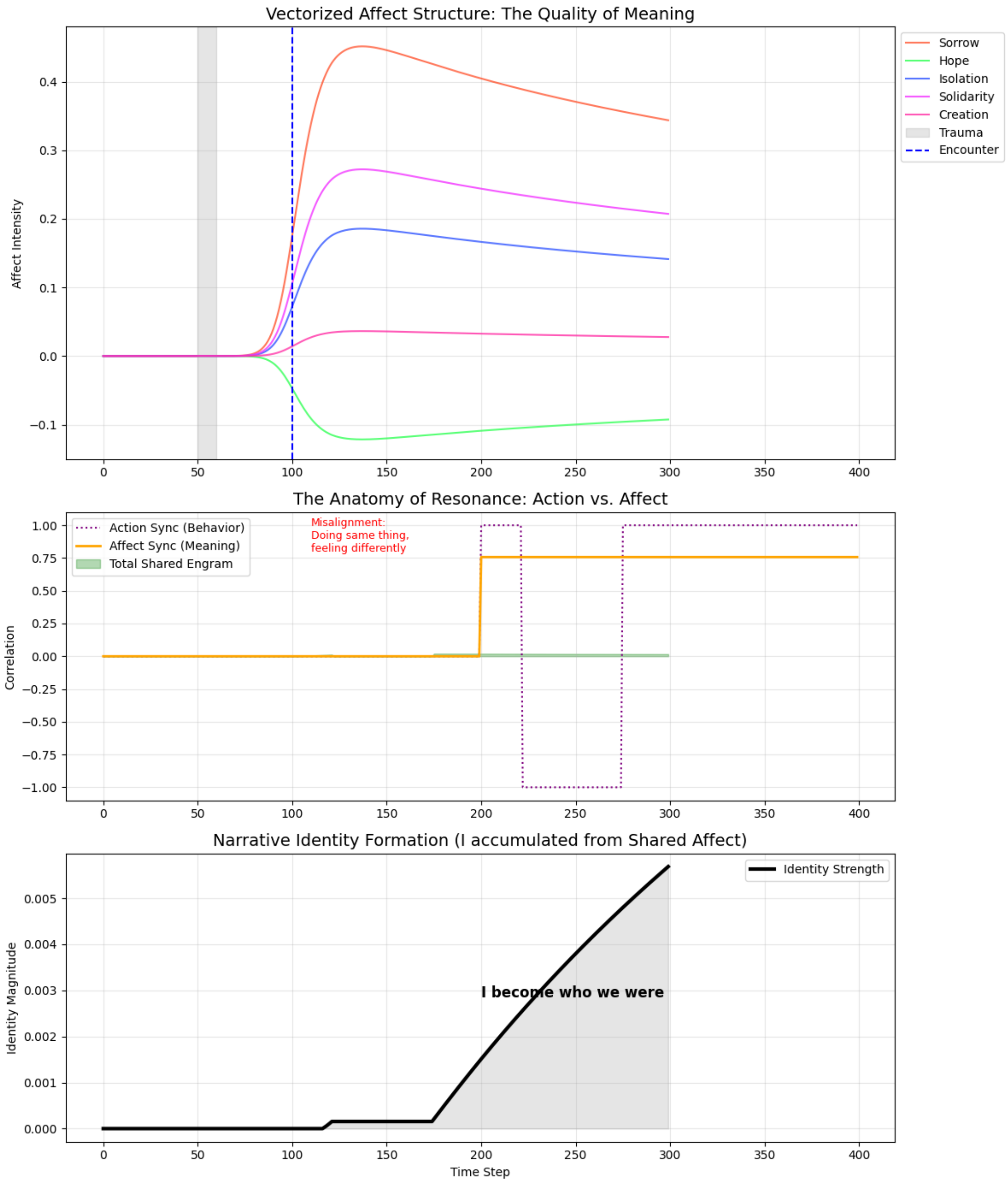

Figure 1: Simulation of Narrative Identity Formation. The Identity (bottom panel) grows only when Shared Resonance (middle panel, green) is active.

3.2 Sensitivity Analysis: The Paradox of Vulnerability

To investigate the role of individual differences, we performed a parameter sweep on Trace Sensitivity α $(0.1 \leq \alpha \leq 2.0)$. As shown in Figure 2, there is a strong positive correlation between α and the final magnitude of Identity.

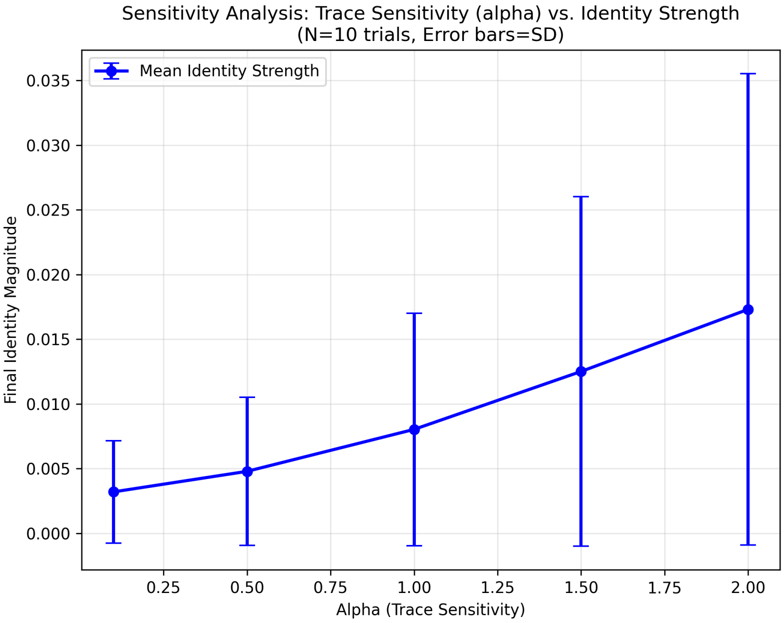

Figure 2: Sensitivity Analysis (N=10 trials). Error bars indicate standard deviation. Agents with higher sensitivity (α) form significantly stronger identities, though with increased variance, suggesting a high-risk, high-reward dynamic.

4 Discussion

The results present a counter-intuitive insight: **Vulnerability is Strength.** Agents that easily ignore traces ($\alpha \approx 0$) minimize immediate costs but fail to accumulate the affective resources necessary for deep social resonance. Conversely, agents that retain traces ($\alpha > 1.0$) suffer more initially but use that suffering as fuel for creative action and identity formation.

This suggests that in the context of General Intelligence, ”forgetting” is not always optimal. The ability to be scarred (Imprint) is the engine of becoming someone (Identity).

Limitations and Future Directions While this study establishes the computational validity of the SIA framework through multi-agent simulations, it has not yet been fitted to empirical human data. Future work will address this by applying the model to behavioral datasets from paradigms such as the Iterated Trust Game [7] or Cyberball tasks [8] to verify its construct validity in clinical populations.

5 Conclusion

We demonstrated that Narrative Identity can be computationally modeled as an integral of shared affective history. The SIA model provides a new mathematical bridge between clinical concepts of trauma and engineering concepts of AGI [1, 2].

References

[1] Thomas Parr and Karl J Friston. Generalised free energy and active inference. *Biological cybernetics*, 113(5):495–513, 2019.

[2] Bessel A Van der Kolk. *The body keeps the score: Brain, mind, and body in the healing of trauma*. Viking, 2014.

[3] Dan P McAdams. The psychology of life stories. *Review of general psychology*, 5(2):100–122, 2001.

[4] Paul Ricoeur. Narrative identity. *Philosophy today*, 35(1):73–81, 1991.

[5] Anil K Seth. Interoceptive inference, emotion, and the embodied self. *Trends in cognitive sciences*, 17(11):565–573, 2013.

[6] Antonio Damasio. *The feeling of what happens: Body and emotion in the making of consciousness*. Harcourt Brace, 1999.

[7] Joyce Berg, John Dickhaut, and Kevin McCabe. Trust, reciprocity, and social history. *Games and economic behavior*, 10(1):122–142, 1995.

[8] Kipling D Williams, Christopher KT Cheung, and Wilma Choi. Cyberostracism: effects of being ignored over the internet. *Journal of personality and social psychology*, 79(5):748, 2000.

Appendix A. Variable Definitions and Representational Spaces

This appendix summarizes the core variables and representational spaces used in the Self-Imprint Attribution (SIA) model. Unlike the main text, no theoretical derivations or update equations are included here. Instead, this section serves as a reference for reproducibility, mathematical clarity, and future implementation.

A.1 Core Variables

Table 1: Core Variables in the SIA Model

Symbol	Psychological Meaning	Computational Role	Mathematical Type
E_t	Emotionally salient experience (event with potential meaning)	External sensory observation at time t	$\mathbb{R}^n$
$\hat{E}_t$	Expected or imagined experience	Prediction from generative model	$\mathbb{R}^n$
D_t	Discrepancy between expected and actual experience	Surprise / prediction error signal	$\mathbb{R}^n$ or $\mathbb{R}$
T_t	Imprinted trace of unresolved experience	Latent pre-affective memory representation	$\mathbb{R}^m$
A_t	Affect vector (qualitative meaning)	Semantic and affective interpretation of experience	$\mathbb{R}^k$
I_t	Narrative Identity (historical accumulation of meaning)	Integrated representation of self over time	$\mathbb{R}^k$
π_t	Policy or creative action	Action distribution based on internal state	$\Delta(\mathbb{R}^a)$
$P(\text{Self}\|E_t)$	Self-attribution confidence	Ownership evaluation for experience	$[0, 1]$

A.2 Representational Spaces

Table 2: Representation Spaces in the SIA Framework

Space	Interpretation
$\mathbb{R}^n$	Sensory / experiential input space
$\mathbb{R}^m$	Trace memory manifold (non-Markovian history)
$\mathbb{R}^k$	Affect and identity representational manifold
$\Delta(\mathbb{R}^a)$	Action-policy distribution space

Appendix B. Simulation Settings and Hyperparameters

This appendix summarizes the experimental settings and hyperparameters used for the simulations presented in the manuscript. These details are not required for theoretical understanding, but are essential for reproducibility and implementation.

B.1 State Dimensionality

Table 3: Dimensionality of Core Variables

Variable	Dimensionality
Sensory input E_t	$n = 6$ (visual, auditory, contextual features)
Trace state T_t	$m = 4$ (latent unresolved imprint)
Affect vector A_t	$k = 6$ (qualitative embedding of meaning)
Identity vector I_t	$k = 6$ (same representational manifold as A_t)
Action / policy π_t	$a = 3$ (expressive, avoidance, reflective)

B.2 Core Hyperparameters

Table 4: Hyperparameter Values for SIA Model Simulation

Parameter	Value
Learning rate η	0.01
Trace sensitivity α	0.25 – 0.40 (varied across runs)
Affect persistence β	0.15
Identity stability S_t (initial)	0.5
Self-attribution threshold	0.6
Simulation length T	200 – 400 timesteps
Random seed	42 (fixed)

B.3 Initial Conditions

- Identity starts from a neutral baseline: $I(0) = \mathbf{0}$.
- Affect is initialized with small random noise: $A(0) \sim \mathcal{N}(0, 0.05)$.
- Trace is initialized as a near-zero memory state: $T(0) = \epsilon \cdot \mathbf{1},\ \epsilon = 0.01$.
- All agents share identical structural models but receive different inputs E_t.

B.4 Update Schedule

1. Compute discrepancy: $D_t = E_t - \hat{E}_t$
2. Update trace: $T_{t+1} = T_t + \gamma \cdot \text{tanh}(\|D_t\|)$
3. Compute attribution probability: $P(\text{Self}|E_t)$
4. Update affect: $A_t = f(T_t, P(\text{Self}|E_t))$
5. Update identity only when $P(\text{Self}|E_t) > 0.6$

Table 5: Evaluation Metrics Used

Metric	Interpretation
Identity variance	Degree of self-structure consolidation
Attribution entropy	Stability of self-boundaries
Affect alignment	Degree of semantic coherence
Trace decay rate	Memory resolution strength
Shared resonance index	Strength of inter-agent synchronization

B.5 Evaluation Metrics

The reported figures (Figures 1–2) were generated using these standardized settings unless otherwise specified.

Appendix C: Simplified PyTorch-style Pseudocode

This section provides a minimal example of the Self-Identity Attribution (SIA) Agent in PyTorch-style pseudocode. For full implementation, see our GitHub repository: https://github.com/rysawaki/Affective_SIA

Listing 1: SIA Agent core loop

```
class SIAAgent:
    def __init__(self, alpha, beta):
        self.trace = torch.zeros(T)        # E_t
        self.affect = torch.zeros(T)       # A_t
        self.identity = torch.zeros(T)     # I_t
        self.engram = torch.zeros((T, D)) # memory of events
        self.alpha = alpha
        self.beta = beta

    def attribution_gate(self, event, identity):
        # Compute P(Self | E)
        sim = F.cosine_similarity(event, identity)
        return torch.sigmoid(self.beta * sim)

    def step(self, e_t):
        p_self = self.attribution_gate(e_t, self.identity)
        self.trace = (1 - self.alpha) * self.trace + self.alpha * e_t
        self.affect += p_self * self.trace
        self.identity = F.normalize(self.affect)
```

This simplified code captures the main mechanisms:

- Event trace accumulation (E_t)
- Attribution via P(Self $\mid E$) using cosine similarity
- Identity update from affective accumulation (A_t)

More details, including multi-agent settings and shared engram computation, are available in the full source code.