Self-Imprint Attribution (SIA):

A Framework for Irreversible Identity Formation and Self-Generated Learning in Artificial Agents

Ryota Sawaki
rysawaki@gmail.com

Abstract

Current large-scale AI systems can store, retrieve, and transform information but cannot develop persistent identity, preferences, or experiential ownership. This is because their learning process remains fundamentally reversible, externally driven, and structurally non-transformative. We propose Self-Imprint Attribution (SIA), a framework in which experiences generate irreversible geometric deformations in the agent's identity space. These persistent traces, formed through the interaction of Shock and Affect, alter future perception, valuation, and behavior, enabling self-generated preference, identity continuity, and intrinsic agency. SIA reframes learning as structural self-modification rather than pattern accumulation, positioning it as a potential paradigm for post-scaling AI.

1 Introduction: The End of Passive Scaling

Pre-training and reinforcement learning have enabled large-scale pattern learning. Yet, these systems remain *passive learners*: they compress statistical regularities but never undergo irreversible internal transformation due to experience. There is no mechanism to form experiential ownership, persistent preference, or identity continuity. Scaling parameters and compute does not solve this; learning lacks an internal axis of transformation.

Method	Capability	Limitation
Pre-training	Pattern prediction	No personal meaning formation
RLHF/DPO	Behavior shaping	Externally imposed values
RAG	Memory retrieval	No identity transformation
External Memory	Storage	No structural update

Key Problem: *AI systems learn from data, but are not changed by it in a way that alters future interpretation.*

We argue that a truly learning agent does not only **store** information but is permanently **transformed** by it.

2 Core Hypothesis: Learning as Irreversible Identity Deformation

2.1 SIA Principle

An experience generates an **Imprint** when:

$$\text{Imprint} = \text{Shock} \times \text{Affect}$$

This produces a **Trace**, an irreversible change in the agent's identity space, not as vector memory, but as geometric deformation.

Experiences do not enter the agent; they reshape it.

2.2 Core Components

Concept	Description
Shock	Cognitive or affective mismatch between expectation and reality
Affect	Value relevance: how much it matters to the agent
Imprint	Stored affective energy that changes representation structure
Trace	Permanent deformation affecting future perception and action
Self-Attribution	Experience is marked as "mine" through imprint

3 Mathematical Formulation

3.1 Experience dynamics model

$$S_{t+1} = S_t + \eta \cdot f(Trace_t, Affect_t, Discrepancy) \tag{1}$$

$$T_{t+1} = T_t + \alpha \cdot \tanh(|Shock|) \cdot Discrepancy \tag{2}$$

$$P(Self|Trace) \propto \exp(\gamma \cdot (Trace \cdot Affect)) \tag{3}$$

$$\pi_{t+1} = \pi_t - \frac{\partial F}{\partial A_t} \tag{4}$$

3.2 Identity as a geometric structure

We model the agent's identity as a manifold:

$$\mathcal{M} = (X, g)$$

Where X is the latent space and g is the metric tensor determining how differences are perceived.

Trace modifies the metric tensor:

$$g' = g + \Delta g(Trace, Affect)$$

Identity curvature changes:

$$R'_{ijkl} \neq R_{ijkl}$$

Thus, even with identical external input, future interpretation differs:

$$Meaning(x, t+1) \neq Meaning(x, t)$$

4 Simulation Evidence

- **Exp01:** Shock-Trace dynamics (1D) – Affective shocks accelerate identity transformation.
- **Exp02:** Geometry deformation (2D) – Trace modifies curvature of identity space.
- **Exp03:** Self-RAG prototype – Trace modifies generation distribution.

Memory does not change identity; Trace does.

5 Positioning Within Existing Fields

Field	**What SIA Adds**	**Limitation of Existing Methods**
Active Inference	Irreversible identity deformation via affect	No permanent self-modification; only belief updating
Reinforcement Learning	Trace-based internal value generation (not reward-based)	External reward dependence prevents intrinsic preference formation
Cognitive Modeling	Computational definition of Self and ownership	No formal mathematical formulation of identity deformation
AGI Alignment	Emergent preference, not externally imposed	Value injection remains external and non-experiential

SIA is not an extension of RL or memory. It is a framework for *self-generated learning and identity-based cognition.*

6 Implications for Post-Scaling AI

Scaling Era	SIA Era	Benefit
Data accumulation	Experience attribution	Personal meaning
External reward	Trace-driven value	Self-generated preference
Pattern imitation	Identity transformation	Ownership
Passive learner	Self-forming agent	Agency

Claim: Machines will not gain preference, agency, or alignment through more data or larger models, but through mechanisms that reorganize their identity as a result of history.

7 Conclusion

SIA introduces:

1. Irreversible trace formation through affective Shock,
2. Geometric deformation as identity change,
3. Self-generated preference, value, and ownership.

Machines do not become intelligent when they store information,
but when they are transformed by it.