Primordial Identity Superposition:

A Computational Framework for Identity, Memory, and Moral Attenuation in Human–AI Systems

Author: Jongmin Lee
Location: Jeonju, South Korea
Areas of Focus: Identity Modeling, Memory Architecture, Moral Attenuation, Human–AI Interaction
Contact: LinkedIn: https://www.linkedin.com/in/jongminlee-schumzt/
Email: zizou9210@gmail.com
Date: 2025.11.24

Abstract

Most AI alignment and personalization frameworks assume that a "user" is a single, coherent, and temporally stable agent. In practice, however, human identity is fragmented, context-dependent, and capable of sudden phase transitions under psychological stress, trauma, or radical life events. This portfolio proposes **Primordial Identity Superposition (PIS)**: a computational framework that models human identity as a *dynamic superposition of latent identity states* rather than a single stable profile.

In the PIS model, a person at time t is represented as a probability distribution over identity states, each with its own goals, memories, moral priors, and agency level. High-stakes events cause **identity collapse**, selecting one dominant state; subsequent experience reforms a new superposition. This structure is integrated with **Moral Attenuation Theory (MAT)**, which models how moral concern decays with psychological distance, time, and abstraction—affecting which identity states remain active or degrade over time.

I argue that next-generation AI systems—especially personal agents, safety layers, and alignment-critical models—should treat users not as static vectors but as **multi-state, phase-shifting systems**. I present:

1. A formal identity superposition model using vector and matrix representations.
2. Collapse and reconstruction dynamics under contextual pressure.
3. A sketch of identity-indexed memory architectures.

4. Alignment and safety implications for OpenAI's products and research agenda.
5. A multi-agent extension to societal identity fields.

This portfolio is both theoretical and experiential: the framework is informed by long-term self-observation under severe psychological stress and subsequent reconstruction, then translated into a computational language suitable for engineering. It is intended as a research proposal, conceptual whitepaper, and collaboration invitation for exploring identity-aware AI alignment.

---

1. Introduction

Most current AI systems, including large language models, implicitly assume that:

- there is *one* user,
- with *one* coherent set of preferences,
- and that those preferences are sufficiently stable to be "aligned to."

But human life does not behave this way. Real people:

- hold contradictory values at the same time,
- act like different persons in different relationships,
- "become a different self" under extreme stress, and
- sometimes rebuild themselves entirely after trauma, success, or collapse.

Traditional philosophical accounts—Locke's memory continuity, Kant's rational agency, Parfit's psychological connectedness—capture parts of this story, but they are rarely expressed in a form that can be used directly in AI engineering.

At the same time, modern AI products (e.g., personal assistants, copilots, therapeutic systems) increasingly shape:

- what users pay attention to,
- how they interpret their past,
- and who they gradually become.

This creates an alignment challenge that is not only about *behavior* but about *identity*.

1.1 Core Thesis

The core thesis of this portfolio is:

A human is not a single agent but a **superposition of identity states** whose weights change over time.
AI that ignores this will misalign to temporary or unstable selves.
AI that models this can instead stabilize and protect human identity.

I call this framework **Primordial Identity Superposition (PIS)** because it describes identity at its most fundamental level—not as a narrative, but as a structured distribution that can be computed, updated, and aligned to.

---

2. Conceptual Background

2.1 Identity as Superposition, Not a Single Thread

Instead of "Who is this person?", PIS asks:
"What are the identity states currently active, and how strongly?"

Key intuitions:

- A person has multiple internally coherent identity states: the analytical self, the fearful self, the caregiver, the competitor, the patient, etc.
- At any moment, these states are present with different intensities.
- Under pressure, one state can dominate and temporarily "become the person."

This is analogous to **superposition** in quantum mechanics: many states exist in parallel until a measurement (event) causes a collapse. Here, the "measurement" is a high-stakes decision, threat, opportunity, or emotional shock.

2.2 Memory as Reconstruction, Not Storage

Neuroscience suggests that memory is not a static record but a **reconstruction process**:

- We do not simply "retrieve" the past; we rebuild it depending on our current state.

- Some memories are **state-dependent**: accessible only when a certain emotional or identity state is active.
- This fits naturally with identity superposition: each state has its *own memory emphasis*.

This motivates **identity-indexed memory architectures** in AI: memory should be aware of **which identity state** it belongs to or was formed under.

2.3 Moral Attenuation Theory (MAT)

Moral Attenuation Theory (a related concept developed by the author) proposes that moral concern is not a constant trait but decays as:

- psychological or physical distance increases,
- time separation grows,
- interaction becomes abstract or mediated,
- and shared identity (common ground) weakens.

In simple proportional form:

- *Moral engagement* ≈ 1 / (*distance to the other* + other attenuation factors)

PIS integrates this by letting **moral attenuation influence identity state weights**. States that feel morally disconnected from others become easier to activate in harmful contexts.

---

3. Formal Model of Primordial Identity Superposition

In this section, I present a simplified mathematical formulation. The purpose is not to claim that human identity can be precisely captured by a few equations, but to give engineers and researchers a **formal starting point**.

3.1 Identity States

Let:

- $S = \{ S_1 , S_2 , \cdots, S_n \}$ be the set of latent identity states.

Each identity state S_i is a structured vector:

- $S_i = (\text{Traits}_i, \text{Goals}_i, \text{MemorySet}_i, \text{Agency}_i, \text{MoralPriors}_i)$

Where:

- $\text{Traits}_i \in \mathbb{R}^k$ — personality / temperament embedding
- $\text{Goals}_i \in \mathbb{R}^m$ — active objective vectors
- $\text{MemorySet}_i \subset M$ — subset of memories most accessible to this state
- $\text{Agency}_i \in [0, 1]$ — capacity of this state to act and take control
- $\text{MoralPriors}_i \in \mathbb{R}^d$ — ethical orientation of this state

3.2 Identity as Weighted Superposition

At time t, the person's identity is:

- $I(t) = \Sigma_i \, w_i(t) \cdot S_i$

where:

- $w_i(t) \geq 0$
- $\Sigma_i \, w_i(t) = 1$

The vector **w(t)** represents the **activation weights** of each state. It is analogous to a probability distribution over "who is currently in charge."

3.3 Identity Weight Dynamics

Weights change over time as new experiences, contexts, and rational pressures appear. We define:

- $w_i(t+1) = f(w_i(t), \Delta M(t), C(t), R(t))$

Where:

- $\Delta M(t)$ — memory updates between t and t+1
- $C(t)$ — contextual factors (environment, relationships, threats, opportunities)
- $R(t)$ — rational coherence pressure (the mind's attempt to reduce internal contradictions)

Interpretation:

- Identity is **not static**; it is an evolving distribution similar to a Bayesian posterior, but defined over identity states rather than simple beliefs.

3.4 Identity Collapse

High-stakes or high-stress events cause **identity collapse**: a transition from a broad superposition to a sharply peaked distribution.

We define:

- **Collapse(event) = argmax_i [$w_i(t) \cdot C(\text{event}, S_i)$]**

Where:

- $C(\text{event}, S_i)$ — compatibility of state S_i with the given event (e.g., survival relevance, emotional resonance, learned habit)

Examples:

- During an exam, the "panic-and-guess" identity may collapse into dominance.
- During a crisis, the "caretaker" state may dominate over the analytical observer.
- During a romantic confession, an emotionally vulnerable state may take over.

After collapse, the dominant state S_k acts as the **active agent**.

3.5 Post-Collapse Reconstruction

Identity does not simply reset to its previous distribution. Collapse itself produces learning and restructuring:

- $I'(t) = \text{Normalize}(S_k + \varepsilon \cdot S_j + \text{NewState}(t))$

Where:

- S_k — state that dominated during collapse
- S_j — state(s) that were strongly affected (e.g., suppressed or reactivated)
- NewState(t) — any newly formed identity component
- ε — small mixing coefficient

Interpretation:

- After trauma, success, or intense relational experience, a person is **not the same superposition** as before.
- Identity has undergone a **phase transition** and now evolves on a modified state-space.

3.6 Integrating Moral Attenuation

We define a simplified moral attenuation function:

- $MA = g(\text{distance}, \text{pressure}, \text{shared\_state}, \text{temporal\_lag})$

Where $MA \geq 0$ is higher when moral concern is weaker (i.e., more attenuation).

We then link MA to identity dynamics:

- $\partial w_i / \partial t \propto -MA + \text{RationalAnchoring}$

Intuition:

- As psychological/moral distance increases, some identity states (e.g., empathetic, prosocial) lose weight, while more self-focused or defensive states gain influence.
- RationalAnchoring represents the stabilizing force of reflection, values, and commitments.

This makes Moral Attenuation not just an ethical concept but a **parameter in identity evolution**.

---

4. Computational Architecture for Identity-Aware AI

To make this theory usable for OpenAI-scale systems, we need an implementable architecture.

4.1 High-Level Components

A potential **identity-aware AI module** includes:

1. **Identity State Estimator**
    - Infers current $w_i(t)$ from user behavior, language patterns, and history.

2. Contextual Stress / Collapse Predictor
    - Estimates whether the current context is likely to cause identity collapse.
3. Identity-Indexed Memory System
    - Stores and retrieves memories conditioned on identity state.
4. Policy Engine
    - Generates AI responses that respect identity distribution and safety constraints.
5. Update Engine
    - Updates estimated weights after interaction.

4.2 Identity-Aware Policy

A simple policy formulation:

- $\text{Policy\_AI}(t) = \Sigma_i \, w_i(t) \cdot \text{Response}(S_i)$

Where:

- $\text{Response}(S_i)$ is the response optimized for that identity state (or cluster of similar states).

The AI does not commit to a single "user profile". Instead, it mixes responses according to the current identity distribution.

4.3 Identity-Indexed Memory

Instead of a single monolithic memory:

- $\text{Memory} = \{ M_1 , M_2 , \cdots, M_n \}$

Each M_i stores experiences and facts most relevant to identity state S_i.

Recall is then:

- Recall(event) = f(current_state, event_tags, similarity(current_state, origin_state))

This can explain why:

- A user "forgets" advice given in a calm state during a panic attack.
- Therapeutic progress vanishes when a different identity state is active.

And it suggests:

- AI should consider **which identity was active when a memory was formed**, and **which is active now**.

4.4 Pseudocode Sketch

Identity Update

```
def update_identity_weights(weights, context, params):
    activations = compute_activation(context, params)
    new_weights = softmax(weights * activations)

    collapsed = max(new_weights) > params["collapse_threshold"]
    if collapsed:
        active_state = argmax(new_weights)
    else:
        active_state = "mixture"

    return new_weights, active_state
```

Identity-Indexed Memory Retrieval

```
def recall_event(event, current_state, params):
    similarity = cosine(event.origin_state_vector,
current_state.vector)
    access_prob = event.tag_weight * similarity

    if access_prob > params["access_threshold"]:
```

```
        return event
    else:
        return None
```

These sketches are deliberately simple. The point is not that they're immediately production-ready, but that they show **how PIS can be turned into code paths**.

---

5. Implications for AI Safety and Alignment

5.1 Alignment Over Identity Distributions

Standard alignment objective:

- minimize Loss(output, user_intent)

PIS-informed objective:

- minimize $\mathbb{E}_i$ [Loss(output, intent(S_i)) · $w_i(t)$]

Where:

- intent(S_i) is the intention of identity state S_i

This reframes alignment as **multi-objective optimization** over multiple internal selves, weighted by their current prominence.

5.2 Misalignment Modes in PIS Terms

Misalignment Type	Cause	Example
State Misalignment	AI aligns to wrong identity state	Encouraging revenge when only the angry self is active
Transition Misalignment	Response triggers harmful identity collapse	Pushing high-pressure decisions in fragile states
Memory	AI assumes global memory; user has	User "forgets" advice given outside

Misalignment Type	Cause	Example
Misalignment	state-dependent memory	their current self
Temporal Misalignment	Preferences drifting; AI assumes stability	Addiction, relapse, sudden ideological shifts

PIS makes these failure modes visible and computable.

5.3 Ethical Guardrails

Identity-aware systems must avoid:

- exploiting vulnerable identity states for engagement or monetization,
- amplifying brief emotional spikes into long-term identity shifts,
- pushing users toward narrow or self-destructive identities.

Instead, AI should:

- detect collapse risk,
- prefer stabilizing responses,
- preserve plurality of identity states (not overfit to one extreme).

---

6. Multi-Agent and Societal Extension

6.1 From Single Identity to Population

If a single person is a superposition of states, then a society is a superposition of identity distributions across many individuals.

Let:

- $\text{Society}(t) = \Sigma_j \, S^{(j)}(t) \cdot P(j)$

Where:

- $S^{(j)}(t)$ — identity distribution of person j
- $P(j)$ — influence or power weight of person j

This allows modeling:

- cultural shifts,
- mass polarization,
- collective trauma or euphoria,
- memetic waves amplified by AI systems.

6.2 Identity Wave Propagation

Identity states can propagate through networks like waves:

- $\Delta S_i = \Sigma_{j \in N(i)} f(S_j - S_i) \cdot W_{ij}$

Where:

- N(i) — neighbors of i
- W_{ij} — influence weights (friends, media, algorithms)

This framework can describe:

- radicalization dynamics,
- cancel culture cascades,
- viral trends that reshape identity.

For OpenAI, this offers:

- a way to think about **large-scale identity impact of deployed systems**, not just individual chat sessions.

---

7. Information-Theoretic View of Identity

7.1 Identity as Compression

We can view identity as:

- $\text{Identity}(t) = \text{Decode}(\text{Compress}(\text{Experience}_{1 \to t}))$

Where:

- Compression = how a person summarizes and encodes all past experiences
- Decompression = how they interpret new events using that encoding

Two people with identical experiences but different compression schemes will **become different identities**. AI interactions can modify this compression process by:

- reinforcing certain narratives,
- suppressing or re-weighting certain memories,
- shifting the "loss function" of identity compression.

7.2 Learning Dynamics

We can sketch identity update as:

- $S(t+1) = S(t) + \eta \cdot \nabla \log P(\text{Experience} \mid S(t))$

Where:

- η is an effective learning rate.

Extreme psychological states can be interpreted as distortions of η:

- Trauma → η very large (huge update from one event)
- Depression → η near zero (no update, stuck identity)
- Mania → η unstable and excessive

An identity-aware AI could aim to **stabilize η**, rather than just controlling content.

---

8. Research Roadmap and Collaboration Proposal

8.1 Phase 1 – Formalization and Critique

- Refine mathematical definitions for PIS and Moral Attenuation.
- Clarify assumptions and boundary conditions.
- Submit a preprint or internal whitepaper for feedback from alignment and cognitive modeling experts.

Deliverables:

- "Primordial Identity Superposition" technical preprint
- Formal notation document
- Example implementations on synthetic or toy data

8.2 Phase 2 – Prototype Implementation

- Implement a prototype identity-aware conversational agent.
- Track an approximate identity distribution for volunteer users over time.
- Measure:
  - stability of identity trajectories,
  - collapse events detected vs self-reports,
  - effect of different response strategies on identity coherence.

8.3 Phase 3 – Integration with Safety and Product Teams

- Integrate identity-aware modules as a safety layer:
  - detect risky transitions,
  - switch to stabilizing mode when needed,
  - audit how long-term interaction impacts users' identity structures.
- Apply to:
  - long-term assistants,
  - therapeutic or coaching models,
  - youth-facing systems.

---

9. Author Background (Brief)

I am a Korean educator and independent theorist with:

- a background in teacher education and long-term work as an English instructor,
- extensive self-observation under intense psychological stress (including near-fragmentation of identity and gradual reconstruction),
- ongoing creation of original conceptual frameworks:
  - Primordial Identity Superposition (PIS)
  - Moral Attenuation Theory (MAT)
  - related ideas on memory architecture and identity collapse.

What I bring is not traditional academic credentials alone, but a rare combination of:

- lived experience of identity collapse and recovery,
- strong verbal and analytical skills,
- the ability to translate phenomenology into computational and mathematical form,
- a deep interest in OpenAI's mission of aligning powerful AI with the full complexity of human beings.

This portfolio is not a finished theory but an invitation:

to co-develop a rigorous, testable, and implementable model of human identity for AI systems that must live with us—not just optimize us.