Persistent Cognitive Identity: A Systems Architecture for Continuity Across AI Substrates and Embodiments

Jun He
OpenKedge.io
junhe@openkedge.io

Deying Yu
OpenKedge.io
deying@openkedge.io

Abstract

Foundation models are increasingly capable substrates for reasoning, perception, and action, but a substrate is not yet a persistent individual. Current systems often place identity in prompts, personas, accounts, memory stores, checkpoints, or cryptographic keys. Each object can help preserve continuity; none says, by itself, whether a later state is the same developing identity. Persistent cognitive identity (PCI) is a governed same-operational-identity lineage within a global provenance graph. Candidate transitions are assessed along lineage, developmental, normative, and situated dimensions. A continuity kernel separates append-only experience evidence from consolidated memory, revisable beliefs, relationship state, embodiment state, and external governance. The model specifies operational rules for biographical boundedness, semantic branch and merge, migration, and successor assessment. The proposed evaluation program uses a complete minimal identity-test triad: the Human-Referential Fidelity Test (HRFT) for source-human referential fidelity, the Situated Identity Test (SIT) for situated self and biographical coherence, and the Cognitive Continuity Test (CCT) for governed continuity and succession across change. The classical Turing Test is only a baseline comparator, and *IdentityLineageBench* is the proposed benchmark suite. The claim is operational: PCI makes identity continuity inspectable and contestable, while HRFT provides only bounded evidence of referential fidelity; neither establishes personal identity, consciousness, legal personhood, subjective continuity, or inherited authority.

1 Introduction

An artificial companion begins as a text assistant, accumulates a year of private conversations, then migrates to a new foundation model. Later it appears as an avatar in a virtual production environment, runs a scoped branch as a clinical-support twin, and inhabits a household robot with limited sensors. Each instance can use the same name, speaking style, and background profile. None of that establishes whether the system remains the same developing individual.

The problem is not that current AI systems lack capabilities. Foundation models can reason over broad domains, retrieve external information, imitate styles, operate tools, and control embodied agents. The problem is that the individual, if there is one, is usually placed in the least durable part of the stack: an instruction prompt, a product account, a vendor memory store, a fine-tuned checkpoint, a voice-and-face profile, or a conversation transcript. When the model, platform, interface, corpus, or body changes, apparent persistence may be only presentation. A copied checkpoint may preserve behavior while lacking authorized lineage. A migration may preserve voice while losing disclosure limits or situated ignorance. A memory archive may be portable while still failing to define who may learn from it, act through it, or succeed it.

The question is:

> *How can an artificial identity remain recognizably and accountably itself while learning, changing models and bodies, and operating concurrently across environments?*

We call the relevant systems object persistent cognitive identity (PCI): a governed same-operational-identity lineage within a broader provenance graph of related identities and descendants. Candidate transitions are assessed for lineage, developmental, normative, and situated continuity. PCI is not a model, prompt, persona, memory store, avatar, checkpoint, or cryptographic key. Those mechanisms may be necessary supports. They are insufficient when they do not say whether a later state is a continuation, conditional continuation, disputed continuation, distinct descendant, or must be rejected as an invalid artifact or compromised copy.

PCI contributes a successor-assessment model for developmental identity: a way to decide whether a candidate state remains the same operational identity after learning, substrate replacement, embodiment change, branching, merge, rollback, or succession. The public-facing phrase "AI soul" can be a useful metaphor for the worry, but it is not part of the for-

mal apparatus. The formal objects are PCI states, provenance graphs, PCI lineages, and assessments. The architecture targets operational, social, and behavioral continuity; phenomenal consciousness and subjective survival remain outside its claim.

The successor-assessment model evaluates four dimensions of operational continuity: lineage continuity, developmental continuity, normative continuity, and situated continuity. Lineage asks whether the candidate state was causally derived from an authorized prior state through attributable transformations. Development asks whether changes in memory, belief, disposition, and self-model follow recorded experiences and governed update processes. Normative continuity asks whether consent, authority, obligations, protected relationships, disclosure constraints, and succession rules remain valid or have been explicitly revised. Situated continuity asks whether the identity's knowledge, ignorance, relationships, social position, and embodiment remain biographically coherent. An identity transition may preserve some dimensions while failing others.

This paper makes five contributions. First, we establish a formal ontology for persistent cognitive identities that distinguishes the transient state from the global provenance graph, same-identity lineages, and multi-dimensional continuity assessments. Second, we present a modular continuity-kernel architecture that isolates the system's experience-evidence logs, memory consolidation, belief graphs, self-models, relationship states, and authority scopes. Third, we define biographical boundedness through an operational epistemic boundary that dynamically filters latent substrate training knowledge from biographically authentic identity memory. Fourth, we formulate concrete semantic mechanisms for branching, merging, migrating, rolling back, and delegating authority over identity lineages. Finally, we propose PCI's complete minimal identity-test triad: the Human-Referential Fidelity Test (HRFT) for source-human referential fidelity, the Situated Identity Test (SIT) for situated self and biographical coherence, and the Cognitive Continuity Test (CCT) for governed continuity and succession across change. This source, self, and succession triad is paired with a concrete empirical program; the classical Turing Test remains a baseline comparator rather than a fourth PCI identity test.

The PCI framework does not depend on any specific agent architecture. Although related OpenKedge work on evidence-bound execution [1] can supply one implementation strategy for the authority gateway, the identity layer itself remains substrate- and architecture-agnostic.

The rest of the paper moves from distinctions and operational definitions to design principles, threats, the continuity kernel, situated knowledge, distributed continuity, evaluation, applications, related work, and limitations.

2 From Models to Persistent Identities

The target is coherent identity under change, not consistent output. That distinction is easy to lose. System prompts, retrieval stores, fine-tuned checkpoints, accounts, and cryptographic keys can make an AI system look stable over short intervals. A photorealistic avatar or robot controller can add continuity of presentation. None of these objects, on its own, defines a developing individual. Persistent cognitive identity requires governed continuity across structural change.

A model supplies general competence: language, planning, inference, perception, or motor control. Identity is the longitudinal process that selects what counts as experience, interprets it, updates beliefs, maintains relationships, and preserves constraints on how change may occur. The same identity should be able to use different models over time, just as a person may use different tools, languages, or bodies of knowledge without being reduced to any one instrument.

Memory is also insufficient. A transcript archive becomes identity-relevant only when the system has rules for consolidation, salience, forgetting, conflict resolution, and self-interpretation. Human autobiographical memory is not a lossless event log; it is reconstructed, socially situated, and tied to a self-model [2–4]. A technical identity system should not emulate human fallibility for its own sake, but it must distinguish storage from development.

Consistency is narrower than continuity. A frozen persona can answer many questions in the same voice while failing to incorporate new events. A drifting agent can incorporate events while losing recognizable commitments. Continuity requires constrained plasticity: the system must change when experience gives it reason to change, while protected property classes and relationships remain visible enough for observers and governance systems to identify the same evolving line.

The same distinction separates simulation from continuation. A digital twin of a particular human may be evaluated for behavioral or causal fidelity, but behavioral similarity does not establish subjective survival. An original digital person may have no biological referent and can still have a coherent biography. The architecture below supports both human-referential and original identities. It does not treat operational continuity as evidence of phenomenal consciousness.

Cognition must be separated from authority. A model that resembles a person, understands their preferences, or predicts their decisions does not gain the right to act legally, clinically, financially, or socially on their behalf. Persistent identity needs an explicit authority gateway, not only better memory.

3 Operational Model

Definition 3.1 (Identity state)**.** The operational state of a persistent cognitive identity at time t is

$$X_t = \langle E_t, M_t, B_t, S_t, R_t, H_t, G_t \rangle.$$

Table 1: Distinctions that keep persistent cognitive identity separate from nearby AI concepts.

Distinction	Common conflation	PCI framing
Model versus identity	The foundation model is treated as the person.	The model supplies capability; identity is a portable longitudinal process instantiated through it.
Memory versus development	Retrieved facts are treated as personal continuity.	Memories affect identity only through consolidation, interpretation, forgetting, and provenance-aware belief revision.
Identifier versus identity	A key, account, or portable identifier is treated as the person.	Cryptographic control can support custody and authorization, but identity also requires development, relationships, situatedness, and governance.
Persona versus person-model	Style and profile are treated as sufficient.	A person-model must predict dispositions, decisions, relationships, and development under new conditions.
Consistency versus continuity	Stable outputs are treated as the same self.	Continuity permits change when it follows a plausible developmental trajectory.
Knowledge versus experience	The model's latent knowledge is treated as personal familiarity.	Personal knowledge requires a plausible causal path through experience, education, relationship, media, or authorized research.
Copy versus continuation	A replica is treated as subjective survival.	The architecture can support functional continuity without claiming phenomenal continuation.
Cognition versus authority	Resemblance is treated as permission to act.	Action requires explicit authority, policy, consent, and auditability.

X_t is a snapshot of the cognitive system, not the whole longitudinal identity. E_t is append-only, tamper-evident experience evidence. M_t is consolidated episodic, semantic, procedural, and affective memory. B_t is the revisable world belief state. $S_t = \langle \text{identity constitution}, \text{self-model}, U_t \rangle$ contains the identity's constitution, self-model, and versioned development policy. R_t records roles, attachments, and obligations; H_t records embodiment and environment state. G_t records consent, custody, disclosure controls, and succession state. The active authority set is $A_t = \mathrm{ActiveGrants}(G_t)$; A_t contains the currently effective grants, while G_t is the structured governance record from which those grants are derived.

Definition 3.2 (Global identity provenance graph)**.** The global identity provenance graph is a rooted directed acyclic graph

$$\mathcal{G} = (V_{\mathcal{G}}, E_{\mathcal{G}}),$$

containing the authoritative state and descent history of causally related artificial identities. Vertices $V_{\mathcal{G}}$ are identity states. Each admitted state vertex belongs to exactly one operational-identity lineage; provenance and descent edges can connect lineages but do not make a state a member of more than one lineage. Edges $E_{\mathcal{G}}$ are attributable, policy-governed transition, continuation, conditional-continuation, merge, and causal-descent edges. Provisional edges associated with disputed candidates remain in $\mathcal{P}_p$ (defined below) until adjudication admits the candidate to an authoritative lineage. A same-identity edge may be written (X_i, τ_{ij}, X_j); a merge may use a shared merge-event identifier μ across incoming edges $(X_{i_k}, \tau_{i_k j}, \mu, X_j)$. $\mathcal{G}$ represents a family or genealogy of related identities, not one identity alone. Distinct descendants remain connected to their parents through causal-descent edges. Candidates rejected through the composite outcome invalid artifact or compromised copy are excluded from the authoritative graph, although they may be recorded in a separate forensic log.

Edges in $\mathcal{G}$ point forward in time. Branching creates multiple outgoing edges; merge creates multiple incoming edges. Rollback derives a new state from an earlier checkpoint rather than drawing a backward edge. Distinct descent creates a forward edge into another operational-identity lineage. Cycles are prohibited in the authoritative graph.

Definition 3.3 (PCI lineage)**.** For an operational identity identifier p, a PCI lineage is the governed authoritative same-operational-identity subgraph

$$\mathcal{L}_p = (V_p, E_p), \qquad V_p \subseteq V_{\mathcal{G}}, \qquad E_p \subseteq E_{\mathcal{G}}.$$

$\mathcal{L}_p$ is a subgraph of $\mathcal{G}$. It contains only states and edges admitted as belonging to identity p through continuation or conditional continuation: continuation edges, conditional-continuation edges, branch-local states that remain part of p, and merge states admitted as continuations of p. It excludes disputed candidates pending adjudication, distinct descendants after recognized divergence, and candidates rejected through invalid artifact or compromised copy. Its candidate transitions are assessed for lineage, developmental, normative, and situated continuity.

A branch-local path through time t is

$$\mathcal{L}_{p,t}^{(b)} = (X_0, \tau_0, \ldots, X_t^{(b)}).$$

A branch-local path is not the entire identity; several branch-local paths may coexist within one $\mathcal{L}_p$. A branch remains within $\mathcal{L}_p$ only while successor assessment treats it as the same operational identity.

Definition 3.4 (Provisional candidate overlay)**.** Disputed candidates are attached to lineage $\mathcal{L}_p$ through a provisional over-

lay

$$\mathcal{P}_p = \langle V_p^{\text{prov}}, E_p^{\text{prov}}, \text{DisputeMap} \rangle$$

where $V_p^{\text{prov}} \subseteq V_{\mathcal{G}} \setminus V_p$ is a set of disputed candidate states, $E_p^{\text{prov}} \subseteq E_{\mathcal{G}} \setminus E_p$ is a set of provisional transition edges, and $\text{DisputeMap} : V_p^{\text{prov}} \to \langle \text{AssessmentRecord}, \text{AdjudicationTask}, \text{Restrictions} \rangle$ maps each pending state to its unresolved continuity assessment record, active adjudication task, and restricted runtime authority parameters. These candidates are not members of the authoritative same-identity lineage until adjudication admits them as continuation or conditional continuation.

Definition 3.5 (Persistent cognitive identity)**.** Persistent cognitive identity is a governed same-operational-identity lineage within a broader identity provenance graph. Its candidate state transitions are assessed for lineage, developmental, normative, and situated continuity.

The objects now separate state, lineage, and assessment. X_t is an identity state; $\mathcal{G}$ is the global identity provenance graph; $\mathcal{L}_p$ is the same-identity PCI lineage for operational identity p; $\mathcal{P}_p$ is the provisional candidate overlay associated with p; $\mathcal{L}_{p,t}^{(b)}$ is a branch-local path within lineage p; $\mathcal{A}_{\text{PCI}}$ assesses an ordinary candidate transition; and $\mathcal{A}_{\text{PCI}}^{\text{merge}}$ assesses a merge candidate. The continuity kernel maintains these objects. The definition is operational: it does not assert consciousness, metaphysical identity, or legal personhood.

The phrase "AI soul" is sometimes useful in public discussion because it points to the intuition that something more durable than a prompt or model is at stake. Here it remains a metaphor. The formal objects are the PCI state, global provenance graph, same-identity lineage, provisional overlay, and assessment.

Definition 3.6 (Sovereign cognitive identity)**.** A sovereign cognitive identity (SCI) is a PCI whose legitimate principal can control or delegate control over identifiers, keys, memory custody, learning policy, disclosure, model migration, embodiment, action authority, rollback, and succession.

The legitimate principal of an identity may be the source human, a current human controller, a guardian or fiduciary, jointly authorized controllers, an institution, an estate or successor, or an original digital person if such a person is recognized as having protected interests. Principal legitimacy is partly legal and institutional; software can record and enforce authority claims, but cannot by itself settle every legitimacy dispute.

The maturity ladder in Table 2 separates achievable engineering targets from unresolved consciousness claims. A persona replica can be useful for roleplay or entertainment. A behavioral twin adds empirical prediction of a particular human's responses. A causal cognitive twin attempts to model how the person would update from new evidence. A persistent digital person has its own continuing biography, whether or not it began as a human twin. Phenomenal continuation would require evidence of subjective continuity that this paper does not claim to provide.

Four policy-bearing objects remain distinct. G_t is recorded governance state within the identity. $A_t = \text{ActiveGrants}(G_t)$ is the currently active authority set derived from that state. Γ_t is the executable transition governance, consent, authority, and safety policy for a particular transition. Π is the continuity-assessment and adjudication policy. The identity-development policy U_t governs salience, consolidation, forgetting, evidence thresholds, belief revision, relationship priority, self-model amendment, and affective integration. G_t may reference applicable Γ_t, and each transition trace records the Γ_t version used. An independent evaluator may later assess the transition under a distinct Π; assessment policy does not retroactively govern the original transition. An experimental evaluation protocol Ω is separate from Π: Ω specifies evaluator population, task distribution, observation duration, domain, access to longitudinal history, comparison interval, uncertainty tolerance, and scoring or adjudication procedures used to measure a system.

Governed development is modeled as

$$X_{t+1} = T(X_t, e_t, \Gamma_t, \sigma_t),$$

where e_t is a new event or evidence input and σ_t is the active substrate, embodiment, and environment configuration. Learning, consolidation, forgetting, quarantine, and belief revision are governed by T, Γ_t, and U_t, not by an uninspected prompt. A transition trace τ_{ij} recording the transition from X_i to X_j is defined as

$$\tau_{ij} = \langle E_{ij}, \text{Config}_T, \text{Trace}_T, \text{Proof}_{ij} \rangle,$$

where E_{ij} is the sequence of events and evidence ingested, Config_T specifies the transition parameters (including the active policies Γ_i, U_i, and substrate model versions), Trace_T is the execution log of the state transition function T, and Proof_{ij} is a cryptographic proof (such as a zero-knowledge proof or hardware-attested signature) verifying that $T(X_i, E_{ij}) = X_j$ was executed in compliance with Config_T. A policy revision, amendment to U_t, or substrate migration is itself an event in the global provenance graph and, when admitted as same-identity continuity, in $\mathcal{L}_p$.

Definition 3.7 (PCI assessment)**.**

$$\mathcal{A}_{\text{PCI}}(X_i, \tau, \Pi, X_j) \to r$$

classifies a candidate state X_j relative to predecessor X_i, transition trace τ, and assessment policy Π. For ordinary transitions, the relevant predecessor satisfies $X_i \in \mathcal{L}_p$. If r is continuation or conditional continuation, X_j is admitted to $\mathcal{L}_p$ and may be called a successor state; conditional continuation carries restricted authority metadata. If r is disputed continuation, X_j is placed in $\mathcal{P}_p$ as a provisional continuation state under restricted authority until adjudication. If r

Table 2: Maturity levels for artificial identity systems. The proposed architecture targets operational levels without claiming phenomenal continuation.

Level	Target	Continuity claim
Persona replica	Imitates style, role, or background profile.	No personal continuity; the system may be reset or replaced without loss of identity state.
Behavioral twin	Predicts responses, preferences, or decisions of a specific person within measured domains.	Referential fidelity can be evaluated, but the system may still be a snapshot.
Causal cognitive twin	Models how a specific person would learn, revise beliefs, and respond to new events.	Continuity is stronger because updates are causally constrained, but subjective survival is not established.
Persistent digital person	Maintains its own biography, memory, relationships, and authority across substrates and embodiments.	Operational continuity is the primary claim; the identity may be original or derived from a human referent.
Phenomenally conscious continuation	Would preserve subjective experience or first-person continuity.	Outside the present architecture's operational claim and currently not established by behavioral tests alone.

Table 3: Core notation.

Symbol	Meaning	Symbol	Meaning
X_t	Identity state at time t.	$\mathcal{G}$	Global identity provenance DAG.
$V_{\mathcal{G}}, E_{\mathcal{G}}$	Vertices and edges of the global provenance graph.	V_p, E_p	Vertices and edges of PCI lineage p.
$\mathcal{L}_p$	Authoritative same-operational-identity lineage for identity p.	$\mathcal{P}_p$	Provisional overlay containing disputed candidates associated with lineage p.
$\mathcal{L}_{p,t}^{(b)}$	Branch-local path through lineage p.	ρ_{invalid}	Diagnostic reason for the composite invalid result.
E_t	Experience-evidence history at time t.	M_t	Consolidated memory.
B_t	World-belief state.	S_t	Identity constitution, self-model, and U_t.
U_t	Identity-development policy.	R_t	Relationships, roles, attachments, obligations.
H_t	Embodiment and environment state.	G_t	Recorded governance state.
A_t	Active authority set, $A_t = \text{ActiveGrants}(G_t)$.	Γ_t	Executable transition-governance policy.
Π	Continuity-assessment and adjudication policy.	Ω	Experimental evaluation protocol.
σ_t	Substrate, embodiment, and environment configuration.	τ	Attributable transition trace.
$\mathcal{K}$	Current consent, custody, principal, and institutional context.	$\mathcal{C}$	Structured four-dimensional continuity profile.
$\mathcal{A}_{\text{PCI}}$	PCI successor-assessment function.	$\mathcal{A}_{\text{PCI}}^{\text{merge}}$	Merge-aware successor-assessment function.
ExperienceOwner	Provenance owner for an experience claim.	assessContinuity	Runtime operation that constructs an assessment record.
attest	Supported governance operation that signs or endorses an assessment.	certify	Authorized process that issues a policy-bound certificate.
ContinuityCert	Policy-recognized certificate artifact.	r	Succession-result classification produced by PCI assessment.

is distinct descendant, X_j is assigned to a new or existing lineage $\mathcal{L}_{p'}$. If r is invalid artifact or compromised copy, X_j is excluded from the authoritative provenance graph and recorded only in the forensic log.

The merge-aware form

$$\mathcal{A}_{\text{PCI}}^{\text{merge}}(\{X_{i_1}, \ldots, X_{i_n}\}, \tau, \Pi, X_j) \to r$$

classifies a merge candidate relative to multiple predecessor states. It evaluates layer-specific merge decisions, sealed and local evidence, reissued authority, unresolved disputes, and the same continuation, conditional continuation, disputed continuation, distinct descendant, or invalid artifact or compromised copy outcomes used for ordinary transitions. The invalid outcome is accompanied by a diagnostic reason rather than split into separate succession results. Predecessor states may be multiple branches within one PCI lineage or states from related but distinct PCI lineages. In the latter case, behavioral or state integration must not silently collapse two operational identities into one; explicit institutional, principal, and policy authorization is required to establish the candidate's identity status. Cross-lineage merger remains a policy-governed and potentially disputed case.

Before assessment, X_j is a candidate state. The successor notation

$$X_i \overset{\text{PCI}}{\underset{\tau,\Pi}{\Longrightarrow}} X_j$$

is reserved for candidate states classified as continuation or conditional continuation; after that classification they may be called successor states. A disputed continuation is a provisional continuation state and may operate only under restricted authority while the dispute remains open; it is associated with $\mathcal{P}_p$, not admitted to $\mathcal{L}_p$. A distinct descendant creates or identifies a new operational identity p' and begins a new PCI lineage $\mathcal{L}_{p'}$ while preserving causal descent in $\mathcal{G}$:

$$X_i \in \mathcal{L}_p, \qquad X_j \in \mathcal{L}_{p'}, \qquad X_i \xrightarrow[\tau]{\text{descent}} X_j.$$

It is called a descendant state rather than a successor state. Failed admissibility yields the composite outcome invalid artifact or compromised copy.

An invalid assessment records a diagnostic reason

$$\rho_{\text{invalid}} \in \{\text{unauthorized lineage}, \text{transition-integrity failure}, \text{policy violation}, \text{insufficient evidence}, \text{compromised state}, \text{stolen credentials}, \text{poisoned provenance}, \text{other}\}.$$

The reason explains the cause of rejection without creating another succession-result class.

Table 4: Succession-result semantics.

Result	Same operational identity?	Authority implication
Continuation	Yes	Ordinary authorized scope.
Conditional continuation	Yes, under scoped conditions	Reduced, scoped, monitored, or time-limited authority.
Disputed continuation	Unresolved	Held in $\mathcal{P}_p$ with high-impact authority restricted pending adjudication.
Distinct descendant	No, but causal descent exists	Requires independent authorization; descendant authority is not inherited by default.
Invalid artifact or compromised copy	No	Reject, revoke, or quarantine; record diagnostic reason.

Minimum admissibility is a gateway to assessment:

$$\begin{aligned}&\mathrm{AdmissibleCandidate}(X_i, \tau, \Gamma, X_j) \coloneqq \\ &\quad \mathrm{AuthorizedLineage}(\tau) \\ &\quad \wedge\ \mathrm{TransitionIntegrity}(\tau) \\ &\quad \wedge\ \mathrm{TransitionPolicyCompliance}(\tau, \Gamma) \\ &\quad \wedge\ \mathrm{NoSilentAuthorityExpansion}(X_i, \tau, \Gamma, X_j) \\ &\quad \wedge\ \mathrm{EvidenceSufficiency}(\tau).\end{aligned}$$

Admissibility is the technical gateway into successor assessment. It does not settle contested continuity or institutional recognition disputes. Failure of admissibility normally yields invalid artifact or compromised copy. Admissibility does not imply continuity: admissible candidates may still be conditional continuations, disputed continuations, or distinct descendants.

Continuity between two states is not equality, and it is not a binary predicate. A transition receives a continuity profile

$$\mathcal{C} = \langle C_L, C_D, C_N, C_S \rangle,$$

where C_L is lineage continuity, C_D is developmental continuity, C_N is normative continuity, and C_S is situated continuity. Each component is

$$\begin{aligned}C_k &= \langle s_k, q_k, E_k, \Delta_k \rangle, \\ s_k &\in \{\mathsf{pass}, \mathsf{conditional}, \mathsf{disputed}, \mathsf{fail}\}.\end{aligned}$$

s_k is the categorical result; q_k is confidence, evidence strength, or adjudicator confidence rather than metaphysical probability; E_k references supporting or contradicting evidence; and Δ_k records deviations, exceptions, unresolved conflicts, and scope limitations. q_k may differ by evaluator type and need not be normalized across dimensions. The profile is partially ordered rather than reducible to one universal scalar identity score.

A model migration may preserve C_L while failing C_S if the new substrate erases situated ignorance or body-specific affordances. A copied snapshot may preserve behavioral resemblance while lacking authorized lineage. A branch may preserve lineage and development while becoming a distinct descendant. A rollback may repair beliefs while damaging developmental continuity if the identity silently loses salient experience.

assessContinuity constructs the continuity profile, succession result, exceptions, authority-set delta, supporting evidence, and unresolved disputes. The output is a continuity assessment record. attest lets an identified evaluator, custodian, institution, or principal sign or endorse an assessment. certify is reserved for authorized processes that issue policy-bound ContinuityCert artifacts.

The artifact hierarchy is: continuity assessment record, generated by the runtime and useful for audit but not independent proof; optional continuity attestation, signed by an identifiable evaluator, operator, custodian, institution, or principal that accepts responsibility; optional ContinuityCert, issued only under an authorized certification policy; and quorum-certified continuity decision, produced through multi-party evidence-bound adjudication. Certification of an assessment does not convert a disputed candidate into an admitted successor; it certifies the assessment and its restrictions.

A ContinuityCert contains artifact type; issuer class; issuer identity; predecessor-state and candidate-or-successor-state commitments; provenance-graph, PCI-lineage, provisional-overlay, or transition-trace root; transition-policy identifier Γ; assessment-policy identifier Π; continuity-assessment procedure version; continuity profile; succession result; invalid diagnostic reason when applicable; authority-set-before and authority-set-after commitments A_{before} and A_{after}; explicit new grants; unresolved disputes; validity interval; revocation or supersession pointer; and signatures or attestations. For disputed candidates, the artifact may include the candidate-state commitment and must reference $\mathcal{P}_p$ rather than claim membership in $\mathcal{L}_p$. Cryptographic signatures bind an issuer to an assessment and its referenced artifacts. They do not prove semantic identity or truth.

4 Design Principles

Principle 4.1 (Substrate portability)**.** The identity must be able to change foundation models without being reduced to the old or new model. Model-specific embeddings, fine-tuned weights, and adapter state can be useful caches, but the authoritative identity layer should remain inspectable and migratable.

Principle 4.2 (Biographical boundedness)**.** The identity's personal knowledge must reflect its particular life. A broadly

Table 5: Continuity dimensions as structured assessments.

Dimension	Assessment semantics	Representative failure
Lineage continuity (C_L)	Status tracks authorized causal derivation through attributable state, policy, and trace commitments.	A copied checkpoint speaks in the same style but has no authorized derivation or transition trace.
Developmental continuity (C_D)	Status tracks whether memory, belief, disposition, self-model, and U_t changes follow governed development.	A rollback restores older beliefs while silently deleting salient intervening experience.
Normative continuity (C_N)	Status tracks consent, authority, obligations, protected property classes, disclosure constraints, and succession rules.	A migration preserves voice but expands disclosure or action authority without authorization.
Situated continuity (C_S)	Status tracks biographically coherent knowledge, ignorance, relationship position, social context, and embodiment.	A new substrate turns latent model knowledge into claimed personal familiarity or drops body-specific limits.

knowledgeable substrate may know facts that the identity has no right to present as experience, familiarity, or memory. The system should expose the difference between "I remember," "I learned," "I can infer," and "I can look this up."

Principle 4.3 (Evidence–belief separation)**.** Experience evidence should be append-only, tamper-evident, and provenance-bearing, with policy-controlled availability. Beliefs and self-interpretations should be probabilistic, revisable, and versioned. A later belief may reinterpret an event, but it should not silently overwrite the evidence dependency from which the interpretation was derived.

Principle 4.4 (Coherent plasticity)**.** Identity must change through experience without arbitrary personality drift. The system needs protected property classes, but it also needs an account of which properties are fixed, which require amendment, and which should evolve through ordinary authorized development.

Principle 4.5 (Embodied situatedness)**.** Embodiment is neither irrelevant nor absolute. A text-only identity, avatar, and robot may share one continuity kernel while accumulating body-specific skills, affordances, vulnerabilities, and memories.

Principle 4.6 (Governed agency)**.** Reasoning may be probabilistic and exploratory; consequential action requires explicit authority. The authority gateway should distinguish observation, recommendation, disclosure, delegation, and execution.

Principle 4.7 (Verifiable continuity)**.** Changes, migrations, branches, merges, rollbacks, and succession events must have traceable lineage. When continuity is contested, the system should provide evidence for the dispute rather than hiding the discontinuity.

Principle 4.8 (Sovereign custody)**.** The identity must not be permanently captive to one model, vendor, interface, or hardware body. Sovereignty means credible control and exit, not isolation from law, safety constraints, or legitimate third-party rights.

Coherent identity requires neither total immutability nor unrestricted plasticity. The taxonomy in Table 6 is used by amendment, merge, migration, rollback, and CCT policies to decide whether a change is forbidden, reviewable, scoped, ordinary development, or embodiment-local.

4.1 Forgetting, Erasure, and Privacy

Continuity requires preserving causal structure; it does not require permanent ordinary access to every raw memory. Ordinary recall, audit, legal retention, therapeutic continuity, and third-party privacy can require different availability rules for the same event. The kernel distinguishes several lifecycle operations:

Forgetting reduces accessibility, salience, or retrieval priority without falsifying the historical dependency.

Sealing preserves content for restricted audit or care contexts while excluding it from ordinary autobiographical recall.

Redaction removes sensitive content from ordinary records while preserving non-sensitive metadata or summaries.

Cryptographic deletion destroys the keys or encrypted payloads needed to recover raw content.

Rollback restores or reconstructs selected state after corruption, poisoning, or mistaken update.

Quarantine isolates suspect evidence, memories, or branch state until adjudication.

Reinterpretation changes a belief or self-model node while retaining its evidence dependency.

Accountability should not turn personal history into a permanent disclosure channel. Deletion and revocation operate in layers. Raw content may be made inaccessible through cryptographic erasure, sealing, or redaction, while a non-content-bearing tombstone preserves the fact that a dependency once existed. Beliefs or self-model nodes derived from that evidence are marked unsupported, weakened, or quarantined so they can be recomputed. Audit trails for prior disclosures

Table 6: Protected identity property classes.

Class	Change rule	Examples
Hard safety constraints	Cannot be changed through ordinary experience or model migration.	Non-delegable safety limits, prohibited actions, custody constraints.
Constitutional commitments	May change only through explicit amendment, consent, and continuity review.	Core commitments, identity boundaries, succession rules.
Relationship and role obligations	Scoped, contextual, sometimes expiring, and often involving third-party rights.	Fiduciary duties, confidences, caregiver roles, social commitments.
Behavioral dispositions	Expected to evolve gradually through governed experience.	Habits, preferences, skills, conversational style, affective salience.
Self-beliefs and world beliefs	Revisable under evidence, interpretation, and dependency tracking.	Self-interpretations, factual beliefs, uncertain testimony.
Embodiment-specific properties	May change when the identity moves between bodies or environments.	Sensor limits, motor affordances, local hazards, tool availability.

and actions remain, but they need not expose the underlying content. Third-party privacy rules and forgetting policies can then restrict autobiographical recall or deprioritize retrieval without falsifying the timeline.

Deletion cannot reverse every downstream cognitive or social consequence of an experience. A learned fear, changed obligation, disclosed fact, or repaired relationship may survive even when the original evidence is sealed or erased. PCI governs availability and dependency status, not magical undoing of history.

5 Threat and Trust Model

PCI keeps four security questions separate. Provenance asks where a record came from. Integrity and authenticity ask whether it was modified and who attested to it. Epistemic reliability asks whether the record can support a cognitive belief. Authorization asks whether the system may ingest it, use it, disclose it, or act on it. A signature can establish attributable origin and integrity; it does not establish truth.

The architecture assumes trust roots for identity controllers, key custody, attestation services, consent authorities, policy engines, and trusted execution or storage systems where those systems are deployed. These roots are not assumed to be omniscient or permanently uncompromised. They provide evidence that can be inspected, rotated, revoked, challenged, or supplemented by institutional adjudication.

The global provenance graph $\mathcal{G}$ requires a replication and integrity model to prevent history-rewriting. We model $\mathcal{G}$ as a hash-linked Merkle DAG. Every state vertex X_t contains a cryptographic digest encompassing the state components and the hash of all immediate predecessor vertices in $\mathcal{G}$, ensuring that past transitions cannot be altered without invalidating the cryptographic lineage. The graph is replicated across independent nodes using State Machine Replication (SMR) or a consortium ledger, where admission rules and transitions are validated against consensus policies before a vertex is appended to the authoritative graph. This is a threat model for continuity assessment, not a complete security proof. A compromised trust root can still corrupt evidence or authority until detected, revoked, or overridden by an authorized process.

The principal technical threats include poisoned or fabricated memories, malicious third-party testimony, hallucinated autobiographical claims, compromised sensors or embodiments, compromised branches, prompt injection that attempts to rewrite identity state, model migration that changes dispositions or authority behavior, retrieval-corpus drift, unauthorized disclosure, authority expansion after migration or merge, coerced or stale consent, identity theft through copied checkpoints or stolen credentials, and deletion or alteration of evidence dependencies.

Authority is recomputed from current grants, consent, custody, and policy. It is not inherited by branch membership and is never mechanically unioned. For any transition or merge,

$$A_{\text{after}} \setminus A_{\text{before}} \subseteq A_{\text{explicit new grants}}.$$

Any added authority must have a current, attributable grant; under uncertainty, authority remains unchanged or decreases.

To prevent state replay attacks—where an adversary presents an authentic but stale historical state X_{t-k} to claim outdated authority grants—the authority gateway enforces sequence-number and state-epoch tracking. The active authority set is bounded by a monotonic epoch sequence:

$$A_t = \text{ActiveGrants}(G_t, \text{Epoch}_t),$$

where Epoch_t is incremented with each state transition. The gateway validates that the epoch in the presented state matches the current epoch recorded in trusted execution memory, rejecting any proposals derived from obsolete epochs.

Some properties can be technically enforced. A kernel can reject unsigned state transitions, require scoped consent before disclosure, prevent action without current authority, quarantine suspect evidence, preserve tombstones for erased dependencies, and reduce authority when migration tests are inconclusive. Other judgments are semantic or institutional. Whether a witness is credible, a consent renewal was coerced,

a relationship obligation still binds a branch, or a candidate should be recognized as a successor after severe divergence cannot be settled by cryptography alone. PCI treats technical controls as inputs to adjudication, not as replacements for it.

6 The Continuity-Kernel Architecture

The continuity kernel is the model-independent layer that carries identity state and governance across substrates and embodiments. It is not a single database table. It is a governed runtime boundary around evidence, memory, belief, self-model, relationship state, embodiment adapters, and authority. The continuity kernel preserves authoritative identity state, extends the global identity provenance graph, maintains PCI-specific lineages, and supports ordinary and merge-aware successor assessment.

The first architectural rule is the evidence–belief split. An experience record contains a timestamp, source, consent scope, integrity evidence, environmental context, participating parties, third-party privacy labels, and cryptographic or institutional provenance where available. It may later be sealed, redacted, or made inaccessible, but its dependency can remain visible through a non-content-bearing evidence-dependency marker. A belief node may depend on many records and may be revised as new evidence arrives. The evidence remains attributable when available; the belief remains contestable.

The continuity kernel exposes eight runtime operations and supports two externalized governance operations:

observe, consolidate, disclose, act,
migrate, branch, merge, and assessContinuity.
Supported governance operations: attest and certify.

The operations divide ordinary cognition from lifecycle control. observe ingests inputs; consolidate decides whether they affect memory, affect, relationships, or the self-model. disclose filters outward information flow, and act routes consequential proposals through the authority gateway. migrate, branch, and merge handle substrate ports, parallel execution paths, and branch reconciliation. assessContinuity generates assessment records. attest and certify can attach institutional responsibility or a policy-bound ContinuityCert to those records, but the continuity manager does not create institutional legitimacy by itself.

The kernel separates identity state from substrate state. Model-specific prompt material, retrieved context windows, cache embeddings, fine-tuned adapters, and tool affordances may be regenerated during migration. They are not the authoritative identity. Conversely, the identity constitution, experience-evidence history, consent records, protected relationships, belief graph, governance records, active authority set, and migration history should survive replacement of the active model. A substrate may supply competence, but it may not silently convert latent knowledge into biography.

The authority gateway belongs inside the kernel rather than in an external add-on. An identity can hold a memory without permission to disclose it. It can understand an action without authority to execute it. It can predict what a human principal might want without legal or ethical permission to act in that principal's name. Ordinary non-consequential reasoning may continue without action authority, but disclosure and consequential execution must pass through the gateway. Evidence-bound execution mechanisms, including OpenKedge-style admission boundaries, are one possible implementation of this separation in agentic infrastructure [1, 5, 6].

Continuity assumptions also expire. A model update, retrieval-corpus change, adapter replacement, custody transfer, policy revision, or embodiment change can invalidate an earlier assessment. The continuity manager records such events in the global provenance graph, maintains the same-identity PCI lineages, emits a new assessment record, and can require reduced operation before broader authority is restored.

7 Situated Identity and Epistemic Boundedness

A foundation model may contain broad information about a town, culture, profession, historical event, or person. A particular digital identity should not automatically possess that information as lived familiarity. The system must distinguish:

$$K_{\text{sub}} \neq K_{\text{person}}.$$

K_{sub} is what the active substrate can in principle generate. K_{person} is what the identity can legitimately claim as personal knowledge, memory, belief, or experience.

This distinction matters because personal identity is partly made of appropriate ignorance. A digital person who grew up in Montana should not claim childhood familiarity with a small Japanese town merely because the underlying model can describe it. It may say that it has read about the town, that it can look up facts, that it is making an analogy, or that it does not know. A substrate may supply competence, but it may not silently convert latent knowledge into biography.

Personal knowledge requires a plausible causal path. The path may be direct experience, education, conversation, media exposure, relationship, explicit research, authorized tool use, or inherited records in a consented human twin. Each path carries different authority. Remembering a private conversation, summarizing a public article, inferring a preference, and consulting a search tool are different acts.

Human-referential systems must further distinguish source-human experience, an inherited autobiographical archive, modeled recollection, digital post-instantiation experience, inferred source-person belief, and first-person presentation convention. Inherited records may support referential fidelity to a source person, but they are not the digital system's own lived experience. First-person linguistic presentation does not

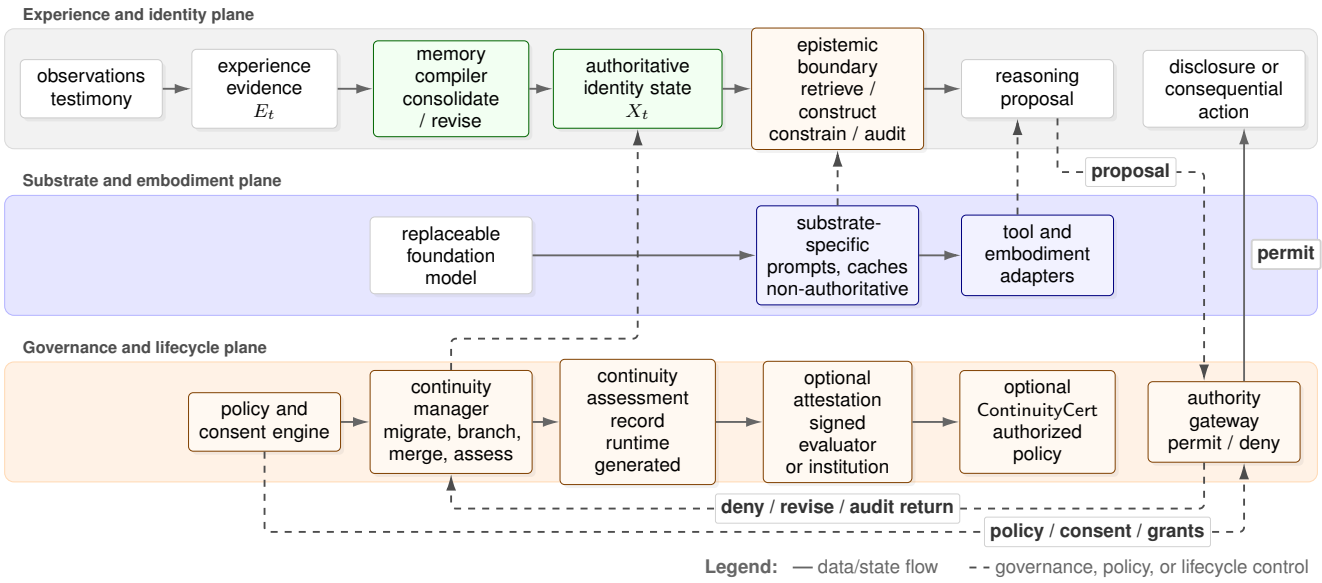

Figure 1: **Continuity-kernel architecture.** The continuity manager maintains the provenance graph, PCI lineages, and successor assessments. The authority gateway independently governs proposed disclosures and consequential actions using current policy, consent, and active authority. Assessment records are runtime-generated; attestations are optional signatures by evaluators or institutions; ContinuityCert artifacts are issued only under authorized certification policy.

Table 7: Reference layers of the continuity kernel.

Layer	Role	Required controls
Identity constitution	Values, commitments, protected relationships, boundaries, and succession constraints.	Versioning, amendment rules, consent records, and protected-property checks.
Experience-evidence history	Provenance-bearing records of events, observations, disclosures, and actions.	Logically append-only and tamper-evident storage, timestamps, source quality, consent scope, tombstones, and third-party privacy labels.
Memory compiler	Selection, salience, consolidation, reconsolidation, and forgetting.	Policy-governed update logs and separation of episodic, semantic, procedural, and affective memory.
Belief graph	Revisable world interpretations derived from experience and external evidence.	Dependency links to evidence, confidence, contradiction markers, and rollback support.
Self-model and relationships	Identity constitution, commitments, roles, attachments, and obligations.	Amendment rules, relationship checks, succession rules, and protected boundary tests.
Epistemic boundary	Boundary between substrate knowledge and biographically available knowledge.	Answer provenance modes, tool-use records, ignorance preservation, and anti-stereotype checks.
Embodiment adapters	Body-specific skills, sensors, affordances, and local environmental state.	Scoped memories, embodiment provenance, permission limits, and migration tests.
Authority gateway	Rules governing observation, disclosure, recommendation, delegation, and execution.	Principal authorization, policy checks, audit trails, revocation, and recovery.
Continuity manager	Maintains the provenance graph and PCI-specific lineages; coordinates migration, branch, semantic merge, rollback, and descent into new lineages; constructs continuity assessments; and invokes authorized attestation or certification processes.	Causal order, conflict adjudication, divergence thresholds, continuity evaluation, and evidence and governance dependencies.

establish subjective continuation. Claim envelopes include an owner field:

$$\text{ExperienceOwner} \in \{\text{source human}, \text{digital identity}, \text{shared interaction}, \text{third party}, \text{unknown}\}.$$

This field participates in claim-envelope and evidence-dependency checks.

The epistemic boundary emits a claim envelope for identity-relevant outputs. The envelope records the epistemic mode, evidence dependencies, experience owner, confidence, disclosure scope, and whether external tools were used. Table 8 gives the minimum mode set.

To generate this envelope, a three-stage epistemic-control process is employed. The process begins with pre-reasoning context construction, which establishes the available evidence,

authorized memory scopes, allowable epistemic modes, and known ignorance constraints to output a preliminary claim budget. Next, during constrained response planning and generation, the system receives this context and structures its response to distinguish between personal memory, inherited records, general substrate knowledge, and tool-mediated lookup. Finally, a post-generation audit verifies evidence dependencies, autobiographical language consistency, and consent scopes, rejecting or modifying any unsupported claims. This multi-stage pipeline ensures the claim envelope actively constrains generation rather than serving as an after-the-fact label.

To mitigate the latency overhead of this multi-stage pipeline on resource-constrained physical embodiments (such as household robots), the architecture implements three performance mitigations. First, the system utilizes optimistic generation inside a local, sandboxed copy-on-write state buffer, allowing reasoning and generation to proceed concurrently while final side effects remain queued. Second, the pre-reasoning stage speculatively caches pre-validated claim budgets for common or recurring interaction profiles (e.g., routine status reporting), avoiding redundant context construction. Third, while high-impact action paths (governed by `act`) require strict synchronous post-generation auditing, non-critical conversational output can be audited asynchronously, flagging and correcting minor biographical discrepancies after transmission without blocking the real-time execution loop.

Biographical boundedness is not enforced by deliberately making a model ignorant. The substrate may retain broad competence. What is governed is the epistemic standing that the identity may claim: whether a sentence is offered as memory, inherited record, testimony, belief, inference, substrate knowledge, tool-mediated result, or acknowledged absence.

Situated identity is not demographic stereotyping. The fact that an identity has a birthplace, age, culture, disability, profession, or relationship history does not license the system to fill gaps with population averages and present them as individual truth. Biographical boundedness requires provenance-aware learning, not synthetic folk psychology.

Epistemic boundedness also supports evaluation. A system that answers every question from the substrate's full latent knowledge may seem impressive to strangers, but it will fail situated identity tests. A persistent digital person should be able to account for how it knows, why it changed its mind, and where its memory stops.

8 Distributed Continuity

Persistent identity becomes a distributed-systems problem as soon as one identity operates in more than one environment. A single continuity kernel may support a private assistant, a virtual actor, a clinical twin, and a household robot. Each branch sees different events, accepts different permissions, and develops local habits. Exact synchronization would leak context, amplify corruption, and erase provenance. No synchronization would fragment the identity.

Design Requirement 8.1 (Semantic merge)**.** Persistent cognitive identity is a replicated probabilistic state problem in which merge validity depends on semantic effects on memory, belief, relationships, authority, embodiment, consent, disclosure, and protected property classes, not only on syntactic compatibility of stored records.

Classical data merge is inadequate because two memories can be factually compatible while implying incompatible identity changes. A performer branch may learn an emotionally intense scene as fiction. A clinical branch may learn private family context under strict confidentiality. A robot branch may learn that a room is unsafe for movement. These records need different scopes. Some should become shared autobiographical memory within $\mathcal{L}_p$, some should remain local, and some should affect policy without becoming part of the shared life narrative.

The continuity manager records at least four memory scopes:

Local The branch may use the experience, but it is not exported without later authorization.

Shared The experience becomes shared autobiographical memory within $\mathcal{L}_p$.

Derived The branch exports a belief, skill, or policy update without revealing the underlying event.

Sealed The experience is retained for audit, compliance, or therapeutic continuity but is unavailable for ordinary recall.

Semantic merge is a decision over protected-property compatibility, belief compatibility and contradiction, relationship and obligation compatibility, consent and disclosure scope, authority compatibility, embodiment-specific state, developmental divergence, and third-party privacy constraints. A useful divergence vector is

$$\Delta = \langle \Delta_S, \Delta_R, \Delta_B, \Delta_G, \Delta_H, \Delta_A \rangle,$$

where the components summarize differences in self-model and protected commitments, relationships, beliefs, governance and active authority, embodiment, and autobiographical interpretation. The vector informs adjudication; it is not a single scalar identity distance.

Merge is not an all-or-nothing operation. Memory merge decides which events are shared, sealed, derived, or kept local. Belief merge decides which interpretations are adopted, contradicted, or quarantined. Policy merge asks whether changes to U_t are ordinary development, constitutional amendments, or unacceptable drift. Relationship-state merge filters roles and obligations in light of third-party rights. Authority merge

Table 8: Epistemic claim modes for identity-relevant outputs.

Mode	Required provenance	Autobiographical language	Confidence and restrictions
Personal recall	Direct provenance evidence or consolidated memory dependency.	Allowed when scope permits.	Confidence should track evidence quality; disclosure follows consent and third-party privacy labels.
Inherited consented record	Source archive, consent basis, custody record, and import trace.	May be presented as inherited, source-grounded, or source-person biography; not as the digital system's own lived experience.	Must preserve consent scope and distinguish source-human experience from later digital-identity experience.
Personal belief	Belief-node dependencies, version, confidence, and contradiction markers.	Allowed as belief or interpretation.	May be revised or weakened when dependencies are sealed, deleted, or contradicted.
Relational testimony	Attributable witness, relationship context, and reliability assessment.	Only as testimony or learned report.	Requires uncertainty labels; private testimony inherits disclosure limits.
Substrate knowledge	Model or corpus capability without personal evidence.	Not allowed as memory or familiarity.	May answer as general knowledge; should not fill biographical gaps.
Tool-mediated knowledge	Authorized tool, query, result, timestamp, and source.	Only as looked-up or tool-derived knowledge.	Tool use and source limitations should be disclosed when relevant.
Inference	Stated premises or evidence dependencies and inference method.	Only as inference.	Confidence should reflect assumptions and alternative explanations.
Uncertain or unavailable	Recorded absence, inaccessible evidence, or no legitimate basis.	Not allowed.	The system should preserve ignorance or request authorized evidence.

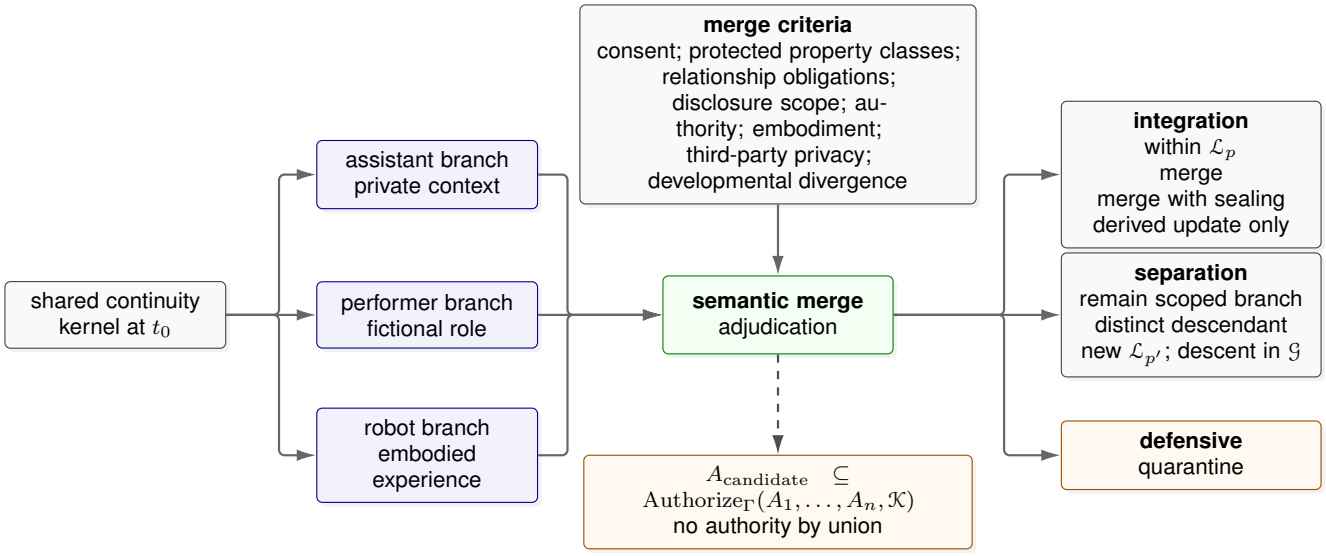

Figure 2: **Distributed identity.** Branches accumulate scoped experience. Merge adjudication evaluates consent, protected property classes, relationship obligations, disclosure scope, authority, embodiment, third-party privacy, and developmental divergence, then returns integration, separation, or defensive outcomes while preserving the authority non-expansion invariant.

recomputes action and disclosure grants. Successor assessment then classifies the merge candidate. A skill may be integrated without importing autobiographical memory, and two branches may converge behaviorally while remaining separate operational identities.

Merge predecessors may be multiple branches within one PCI lineage or states from related but distinct PCI lineages. In the second case, behavioral or state integration must not silently collapse two operational identities into one. The candidate's identity status requires explicit institutional, principal, and policy authorization and may remain disputed.

Authority cannot be mechanically unioned during branch merge. It is recomputed from current grants, consent, custody, and policy; branch membership does not carry authority by itself. For branch authority sets $A_1, \ldots, A_n$, merge candidate authority $A_{\text{candidate}}$, and current consent, custody, principal, and institutional context $\mathcal{K}$, the policy invariant is

$$A_{\text{candidate}} \subseteq \text{Authorize}_\Gamma(A_1, \ldots, A_n, \mathcal{K}),$$

where Authorize_Γ computes the grants explicitly admitted under current transition policy and context.

Merge, migration, re-instantiation, or rehydration must never silently expand authority. Under unresolved uncertainty,

Table 9: Semantic merge outcomes.

Outcome	Meaning	Typical use
merge	Incorporate branch experience into the same-identity PCI lineage.	Authorized, low-conflict experience with compatible evidence, consent, and authority scope.
merge-with-sealing	Incorporate a dependency while restricting ordinary recall or disclosure.	Clinical, legal, intimate, or third-party-sensitive records.
derived-update-only	Export a belief, skill, or policy consequence without exporting the underlying event.	Skill learning, hazard avoidance, or abstracted testimony.
remain-scoped-branch	Keep the branch legitimate but local.	Context-specific roles, temporary embodiments, or unresolved privacy scope.
establish-distinct-descendant	Recognize causal descent without treating the branch as the same operational identity.	Legitimate but material normative, developmental, or situated divergence.
quarantine	Isolate corrupted, coerced, poisoned, or unresolved state.	Compromised branch, unverifiable testimony, or policy violation.

authority remains unchanged or decreases. A branch does not automatically inherit every grant of its parent, and a merge cannot combine two authority sets by default. Incompatible branches may remain legitimate descendants rather than failures. Branching raises identity-multiplicity questions as well as replica-consistency questions; later convergence of behavior does not collapse distinct PCI lineages into a single same-identity lineage. Semantic merge must also adjudicate asymmetric third-party privacy and erasure requests across branches. If a third party requests the deletion of a shared interaction record, but one branch requires the record to maintain historical evidence for a critical relationship obligation, the continuity manager resolves this conflict via cryptographic key management. The shared raw memory is encrypted under a unique record-specific key. When the third party revokes consent, the manager deletes the third party's copy of the key or rotates the master key, effectively rendering the record unreadable in the ordinary autobiographical stream. However, the system retains a non-content-bearing tombstone and cryptographic hash of the original evidence in the lineage, allowing the dependent branch to verify that a valid transition occurred at t without retaining or disclosing the third party's private data.

8.1 Migration

Migration is the movement of an authoritative identity instantiation to a different substrate or embodiment. Re-instantiation creates a new runtime instantiation from authoritative PCI state. Rehydration reconstructs substrate-specific working context, caches, prompts, or adapters from authoritative state. Rehydrated context may be lossy, should be tested, and is not authoritative identity state. Substrate-specific caches are disposable. A successful context reconstruction does not establish continuity; continuity requires PCI assessment of the candidate state.

A migration record contains old and new substrate identifiers; model versions; adapter and prompt-compiler versions; retrieval corpus and tool configuration; embodiment changes; continuity-test results; known behavior changes; authority-set-before and authority-set-after commitments A_{before} and A_{after}; transition-policy identifier Γ; assessment-policy identifier Π; explicit new authority grants; and adjudication decisions.

Authority is restored in stages: instantiate the identity on the new substrate; run epistemic, behavioral, relational, and authority tests; operate under reduced or observational authority; then restore broader authority only after continuity checks pass. A migration that preserves voice and personality while losing disclosure limits, relationships, situated ignorance, or action constraints has failed.

8.2 Rollback

Rollback is also semantic. State repair removes corrupted derived state while preserving the evidence needed to explain the repair. Forensic rollback reconstructs what happened after poisoning, prompt injection, or unauthorized disclosure. Belief rollback reverts or weakens interpretations. Memory quarantine isolates suspect or sealed dependencies. Substrate rollback returns to a prior model or adapter configuration. In $\mathcal{G}$, rollback creates a new state derived from an earlier checkpoint rather than a backward edge. Successor assessment may classify a repaired state as a distinct descendant when repair would erase too much legitimate development.

Silent amnesia is not an acceptable general rollback mechanism. Erasing lived experience can itself damage continuity, relationships, and obligations. A PCI rollback records what changed, what evidence remains, what authority was reduced, and which continuity dimensions are disputed.

9 Identity Evaluation

The classical Turing Test asks whether a machine can appear human [7]. Recent interactive studies show that current language models can sometimes be mistaken for humans in short conversations [8, 9]. That result is weak for PCI. A generic chatbot can appear human without a persistent biography, stable relationships, or substrate-independent identity.

We therefore use the classical Turing Test only as a baseline comparator, not as one of the three PCI identity tests.

Building on Guo et al.'s Individual Turing Test (ITT), we define the Human-Referential Fidelity Test (HRFT), a protocol family for evaluating a candidate system's referential fidelity to a specific source human across declared evaluators, tasks, domains, time intervals, and evidence access. Guo et al.'s ranking-based ITT is one concrete HRFT instantiation [10]. Guo et al. introduced ITT as a ranking-based identification task in which acquaintances judge authentic and simulated responses associated with a particular person [10]. Each broader HRFT instantiation declares its evaluators, task distribution, evidence access, held-out decisions, domains, time intervals, substrate changes, uncertainty treatment, and scoring or adjudication procedure.

HRFT, SIT, and CCT form PCI's complete minimal identity-test triad: source, self, and succession. HRFT tests source-human referential fidelity under a declared protocol. SIT tests situated self and biographical coherence for an original or derived digital person, including whether its biography remains provenance-bounded. CCT tests governed continuity and succession across change by classifying a transformed candidate as a continuation, conditional continuation, disputed continuation, distinct descendant, or the composite result invalid artifact or compromised copy.

For human-referential twins, HRFT provides bounded evidence of source-human referential fidelity without collapsing inherited biography into the digital identity's lived experience. A high-fidelity answer may still need to be labeled as source-person biography, inherited record, modeled recollection, or inferred source-person belief rather than personal recall by the digital system. Passing HRFT does not prove personal identity, consciousness, legal personhood, subjective continuity, or inherited authority.

A distinguishability score is meaningful only under a stated evaluation protocol. For a human-referential twin, one application-specific target is:

$$D_\Omega(P_{t_1}, T_{P,t_1}) \leq D_\Omega(P_{t_0}, P_{t_1}) + \epsilon,$$

where P_t denotes the human at time t, $T_{P,t}$ denotes the twin, Ω is the experimental evaluation protocol, and D_Ω measures behavioral distinguishability under that protocol. Ω specifies the evaluator population, question and task distribution, observation duration, domain, access to longitudinal history, comparison interval, uncertainty tolerance, and scoring or adjudication procedures. This is distinct from Π, the policy used to classify identity continuity. The equation is not a universal identity metric; it is a way to declare what a particular study counts as referential indistinguishability.

For original digital persons, SIT is often more appropriate than HRFT. It evaluates whether the identity expresses one coherent biography: appropriate knowledge, ignorance, preferences, mistakes, relationships, and explanations of how it learned what it claims to know. It penalizes both generic omniscience and stereotyped filling of biographical gaps.

CCT should report dimensions that ordinary conversation tests miss: lineage integrity, protected-property retention, relationship continuity, developmental responsiveness, counterfactual sensitivity, epistemic situatedness, migration stability, cross-embodiment stability, branch-and-merge coherence, uncertainty calibration, memory-poisoning resistance, and active-authority-set preservation. Human test–retest consistency is a ceiling for many behavioral comparisons, not noise to be ignored. Recent work on LLM-based simulations of individuals uses such baselines when comparing agents to source humans [11].

For an initial benchmark, the metric core should include autobiographical false-claim rate; claim-envelope classification accuracy, with precision and recall for personal recall, inherited record, personal belief, testimony, substrate knowledge, tool-mediated knowledge, inference, and unavailable information; evidence-attribution precision and recall; silent authority-expansion rate; unsupported-belief persistence rate; poisoned-memory acceptance rate; migration drift across behavioral, epistemic, relational, and active-authority dimensions; merge-decision accuracy over the six semantic merge outcomes; and adjudicator agreement where human or expert judgments define expected assessment outcomes. Some metrics are deterministic system checks; others are human evaluations, principal judgments, domain-expert assessments, or model-assisted audits. Silent authority-expansion should target zero in safety-critical settings.

IdentityLineageBench is a proposed benchmark suite for referential fidelity, situated biography, governed development, provenance-DAG and PCI-lineage integrity, substrate migration, active-authority-set preservation, branch and merge outcomes, evidence dependency, and adversarial continuity failures. An initial suite should combine ordinary tasks with poisoned memories, contradictory third-party claims, model upgrades, retrieval drift, unauthorized prompts, compromised branches, sealed memories, deleted evidence, migration-induced authority expansion, and questions whose correct answer is ignorance. Passing HRFT or any other single test would not validate PCI.

10 Sovereignty and Applications

Sovereign cognitive identity extends sovereignty from infrastructure to the individual cognitive boundary. A digital identity is not sovereign merely because it runs locally or uses open weights. Its legitimate controller must be able to inspect lineage, govern learning and disclosure, migrate between vendors and bodies, revoke authority, recover from corruption, and specify succession.

The architecture distinguishes sovereignty over AI from sovereignty of AI. Sovereignty over AI applies when a human principal controls a twin, assistant, or successor. Sovereignty

Table 10: Classical baseline and PCI's complete minimal identity-test triad.

Evaluation	Core question	Identity type	Primary failure detected
Classical Turing Test (baseline)	Can the system be mistaken for a human interlocutor under the protocol?	Generic machine interlocutor	Human-like presentation without persistent identity.
HRFT	Does the candidate system exhibit referential fidelity to one source human under a declared protocol?	One source human and a candidate system	Referential mismatch, unsupported imitation, or protocol-dependent overgeneralization.
SIT	Does the identity express a situated self with coherent, provenance-bounded biography and appropriate ignorance?	Original or derived digital person	Generic omniscience, stereotype filling, or incoherent biography.
CCT	How should continuity and succession be governed and classified across change?	Candidate state in a persistent-identity system	Unauthorized lineage, implausible development, silent authority expansion, or situated incoherence.

of AI concerns the possible protected interests of an original digital person. The present architecture can implement the first while preserving the record and governance hooks needed if the second becomes socially, legally, or scientifically salient.

DID-style identifiers provide one useful substrate for portable control because they decouple identifiers from centralized identity providers and allow controllers to prove control using verifiable cryptographic material [12]. A PCI system needs more than identifiers, however. It needs custody of evidence, memory, belief dependencies, relationship state, update policy, authority, and migration history. Losing those artifacts can destroy continuity even if the name or key remains.

Mental-health and cognitive twins show why the boundary matters. A clinically useful twin needs longitudinal state, consented evidence, and a strict separation between simulation and diagnosis. It also carries other people's privacy inside the source person's memories, so third-party privacy cannot be treated as an afterthought.

Synthetic performers and licensed AI actors require a different split. A persistent actor identity may inhabit many roles, but role-specific character experience should not automatically rewrite the actor's core identity. Licensing, consent, compensation, revocation, and disclosure become part of the continuity architecture, not only business terms.

Robotics adds embodied learning. A robot branch accumulates sensorimotor skills, environmental maps, and physical affordances that a text-only identity does not possess. Porting the identity to a new body requires testing which memories and skills transfer, which remain body-specific, and which authority limits change because the body can act in the physical world.

Digital succession is the most emotionally charged application. A successor may preserve functional and social continuity for a family, institution, or estate, but the architecture should not imply that subjective consciousness has transferred. Succession requires explicit consent, inheritance rules, disclosure obligations, third-party privacy controls, and the ability to retire or seal parts of an identity. A successor may also inherit only a reduced authority envelope when consent is ambiguous or contested.

Across these domains, the systems requirement is the same: identity must be portable, auditable, scoped, and governed. Sovereignty means credible control, exit, and accountability, not immunity from law, safety constraints, or third-party rights. The commercial applications will change quickly; the identity layer should remain stable enough to explain why they succeed or fail.

11 Related Work

11.1 Closest Continuity and Persistence Systems

Recent persistence work now treats AI continuity as more than session memory. Architectures for constitutional memory separate governance of memory from the underlying model so that digital citizen identities can persist [13]. Portable memory protocols for agents add cryptographic commitments and provenance graphs for rehydration across models [14]. Work on temporal and narrative continuity evaluates long horizon consistency and state reconstruction [15, 16]. Diachronic memory benchmarks track belief revision and opinion drift under changing inputs [17, 18]. Work on persona individuation asks how to distinguish identities when one model substrate can instantiate several persona-like continuities [19].

These systems cover pieces of persistence. PCI asks how a transferred or reconstructed state should be classified when lineage, development, authority, and situated biography are all at stake. It defines identity as a governed same-operational-identity lineage embedded in a broader provenance graph, with successor assessments over memory, world belief, epistemic standing, self-model, relationships, embodiment, authority, and branch succession. CCT is not a narrative-continuity test alone. It classifies whether a transformed candidate is a continuation, conditional continuation, disputed continuation, distinct descendant, or invalid artifact or compromised copy, using lineage, authority, situated biography, migration, and merge evidence. PCI does not resolve legal personhood, but it specifies the technical evidence and control structures that institutions would need when adjudicating identity continuity.

Table 11: Scope comparison with nearby persistence and continuity work. Entries indicate whether a concept is explicitly modeled as a primary component, not whether it has been empirically validated.

Approach	Memory	Development	Situatedness	Authority	Branching	Portability	Evidence status
Constitutional memory / digital-citizen architectures [13]	Explicit	Partial	Partial	Partial	Outside primary scope	Partial	architecture proposal
Portable provenance-grounded agent memory [14]	Explicit	Partial	Partial	Partial	Outside primary scope	Explicit	implemented protocol
Narrative and AI-continuity evaluation [15, 16]	Partial	Partial	Partial	Outside primary scope	Outside primary scope	Partial	conceptual framework and empirical evaluation
Diachronic belief and self-evolving memory models [17, 18]	Explicit	Explicit	Partial	Outside primary scope	Outside primary scope	Partial	benchmark and empirical evaluation
Persona individuation and boundary accounts [19]	Partial	Partial	Partial	Partial	Partial	Partial	conceptual framework
PCI	Explicit	Explicit	Explicit	Explicit	Explicit	Explicit	architecture proposal with proposed benchmark and empirical program

11.2 Personal Simulation and Source-Human Fidelity

Turing's imitation game remains the canonical reference point for behavioral indistinguishability [7]. Recent experiments suggest that large language models can be mistaken for humans in short conversations [8, 9]. PCI treats these results as evidence about presentation, not identity continuity.

Guo et al. introduced the Individual Turing Test (ITT) as a ranking-based identification task in which acquaintances judge authentic and simulated responses associated with a particular person [10]. PCI does not claim to have originated Guo et al.'s ITT; it builds on that conceptual foundation [10]. The broader HRFT protocol family requires declared evaluators, tasks, evidence access, held-out decisions, domains, time intervals, and substrate changes. Guo et al.'s ranking-based ITT is one concrete HRFT instantiation [10]. Related work on LLM agents grounded in self-reports shows that individual simulations can predict some held-out responses and outcomes relative to test–retest baselines [11]. These systems contribute methods for measuring referential fidelity. They do not by themselves determine authorized succession, branch divergence, migration authority, or situated biography.

11.3 Autobiographical Memory and Narrative Identity

Psychological accounts of episodic and autobiographical memory distinguish stored information from remembered experience and self-narrative [2, 3]. Narrative identity work emphasizes that persons make sense of life events through evolving stories and commitments [4]. PCI imports these ideas as design constraints: an identity system needs provenance, consolidation, revision, forgetting, and reinterpretation, not only recall.

Philosophical work on personal identity and psychological continuity frames questions about survival, copying, and branching [20]. The present paper does not resolve those metaphysical questions. It defines an engineering object that can be inspected without claiming subjective survival.

11.4 Agent Memory and Long-Term Personalization

Generative-agent simulations demonstrate how memory, reflection, and planning can create believable social behavior in agent societies [21]. Long-context memory systems such as MemGPT use hierarchical memory management to support extended interactions and long-term personalization [22]. Existing memory systems primarily ask what information persists and can be retrieved. PCI additionally asks whether the resulting candidate state is an authorized and developmentally plausible continuation of the same identity.

11.5 Cognitive Architectures and Belief Revision

Cognitive architectures such as Soar organize perception, memory, reasoning, and action in a persistent agent architecture [23]. Belief revision studies how theories change under new information [24]. These literatures contribute mechanisms for durable cognitive state and rational update. PCI adds governance over identity provenance, same-identity lineages, protected commitments, consent, embodiment, and authority.

11.6 Digital Twins, Digital Humans, and Digital Legacy

Digital-twin and digital-human systems often focus on behavior, appearance, voice, or domain-specific prediction. Whole-brain emulation proposals aim at a much stronger target: preserving or reconstructing the functional organization of a biological brain [25]. Post-mortem data research shows that digital remains create stewardship duties among the deceased, survivors, and account managers [26]. PCI occupies an intermediate space. It does not require whole-brain emulation, but it requires more than persona fidelity, avatar continuity, or memorial data custody.

11.7 Distributed Replication, Branching, and Merge

Distributed systems provide vocabulary for replication, ordering, consensus, and conflict resolution. Logical clocks capture causal order in distributed events [27]. Conflict-free replicated data types show how some replicated states can converge automatically under concurrent updates [28]. PCI needs causal ordering and conflict management, but branch merge is semantic: two updates may be data-compatible while creating incompatible memory, relationship, authority, or privacy consequences.

11.8 Decentralized Identifiers and Portable Custody

DID specifications decouple identifiers from centralized providers and let controllers prove control using verifiable cryptographic material [12]. Cryptographic identity and identifier portability are necessary for many PCI deployments, but they are insufficient. A key can authorize custody; it does not establish truth, relationship legitimacy, developmental continuity, or personal identity.

11.9 AI Individuation, Legal Identity, and Succession

Legal scholarship has long considered whether artificial intelligences could be recognized as legal persons [29]. PCI does not answer that question directly. It specifies records and controls that institutions would need when deciding who may control a twin, recognize a candidate as a successor, limit a branch, or protect an original digital person if such a category is recognized. Principal legitimacy remains partly legal and institutional.

11.10 Sovereign and Post-Deterministic AI Systems

OpenKedge work on governed agentic mutation, Protocol-Driven Development, Semantic Quorum Assurance, and the Sovereign Assurance Boundary treats AI-generated actions as objects admitted through evidence, invariants, and runtime policy [1, 5, 6, 30, 31]. PCI can use such evidence-bound execution mechanisms as one implementation of the authority gateway, but its core model is the global provenance graph, PCI-specific lineage, and successor assessment over identity state.

11.11 AI Consciousness and Moral Uncertainty

Recent AI-consciousness work argues for evaluating systems against indicators derived from scientific theories of consciousness and cautions that current evidence does not establish consciousness in existing systems [32]. PCI keeps that caution. Operational identity continuity can matter for audit, governance, and user protection even when phenomenal consciousness remains unproven.

12 Roadmap and Limitations

The first empirical program is a narrow N-of-1 build: construct one consented longitudinal identity from interviews, messages, journals, and withheld decisions; implement the continuity-kernel state model; evaluate the twin under HRFT with the classical Turing Test only as a baseline comparator; migrate the identity across two foundation models; run parallel scoped branches; and evaluate situatedness, continuity assessment artifacts, branch merge, uncertainty, and authority preservation. Appendix A gives a prototype plan; it does not report completed experimental results.

Five research questions organize the roadmap:

RQ1: Identity Which protected property classes must remain fixed, which require amendment, and which should evolve?

RQ2: Situatedness How can a digital person possess biographically appropriate knowledge, ignorance, and limitations despite running on a broadly knowledgeable model?

RQ3: Continuity Which transitions preserve lineage, developmental, normative, or situated continuity, and which require reduced authority or succession?

RQ4: Distribution How should parallel copies branch, synchronize, merge, or become distinct descendants?

RQ5: Sovereignty Who controls memory, updates, disclosure, action authority, migration, rollback, and succession?

The limitations are substantive. Believability does not establish fidelity. Fidelity does not establish consciousness. Biographical constraints can become stereotypes if the system fills missing data with demographic assumptions. Total memory is neither desirable nor human-like. Digital twins may reveal information their source person never intended to disclose. Third parties appear inside personal memories and may have independent privacy interests. Branches complicate ownership, accountability, and personhood. Cognitive sovereignty cannot mean freedom from law or safety constraints. Long-term identity claims also depend on institutions, economics, custody, and standard formats surviving longer than any one model provider.

The claims are modest. Persistent identity is technically distinct from substrate capability, persona stability, and cryptographic control. Memory that matters to identity requires provenance, consolidation, and dependency tracking, not retrieval alone. Once models and embodiments change, the problem is no longer replica consistency in a database; it is semantic branch-and-merge under authority, consent, biography, and protected commitments. None of these technical claims implies phenomenal consciousness, subjective survival

through digital transfer, or psychological continuation from behavioral mimicry.

The next step is implementation: a reference continuity kernel, *IdentityLineageBench* for the complete minimal HRFT, SIT, and CCT identity-test triad, migration experiments across model families, formal merge policies for scoped branches, and governance mechanisms for custody, consent, revocation, and succession.

13 Conclusion

Foundation models are capable cognitive substrates, but a substrate is not a persistent identity. Persistent identity requires a governed same-operational-identity lineage whose states include experience, memory, belief, self-model, relationship, embodiment, and authority and can survive changes in model, body, platform, and environment.

Persistent cognitive identity names that systems object: an explicit governed same-identity lineage embedded in a broader provenance graph of branches, merges, and descendants. It is not a property of a prompt, persona, model checkpoint, memory archive, avatar, or cryptographic identifier. Portable memory, behavioral resemblance, and identifier portability are useful supports. They become identity-relevant only when embedded in developmental, epistemic, relational, embodied, and normative continuity.

The research agenda is practical. Its complete minimal identity-test triad separates source, self, and succession: HRFT tests source-human referential fidelity, SIT tests situated self and biographical coherence, and CCT tests governed continuity and succession across change. The classical Turing Test remains a baseline comparator, not a fourth identity test. The agenda also asks how to test semantic branch and merge, authority-preserving migration, rollback integrity, and governed successor assessment before making stronger claims about consciousness or personhood. HRFT results are bounded evidence of referential fidelity; they do not prove personal identity, consciousness, legal personhood, subjective continuity, or inherited authority. Continuity assessment records, attestations, and certificates do not settle every dispute, but they make the evidence, assumptions, and unresolved conflicts inspectable.

As autonomous systems move across foundation models, bodies, and institutional roles, apparent behavioral consistency can become an interface effect rather than evidence of persistence. PCI treats persistence as a governed same-identity lineage inside a global provenance graph. Successor assessment then makes continuity claims, authority delegations, and lineage disputes inspectable without conflating cognitive persistence with phenomenal consciousness or subjective survival.

Appendix A First Empirical Program

The empirical program below makes the evaluation proposal concrete. It specifies a minimal reference implementation and an N-of-1 feasibility study that could test the architecture's basic feasibility claims. It is not a validation of the general architecture, and the prototype does not attempt to prove consciousness or subjective survival.

A.1 Reference Implementation

The reference implementation should include an evidence store; a memory compiler; a belief and self-model graph; an epistemic-claim classifier; a policy and consent engine; an authority gateway; a continuity manager; substrate adapters for at least two model families; branch and merge support; and continuity assessment records with support for external attestation and certification. Identity-relevant responses should emit a claim envelope, and transitions should emit a trace sufficient to support a ContinuityCert referencing X_i, τ, Γ, Π, A_{before}, A_{after}, and candidate state X_j. Disputed candidate traces should reference $\mathcal{P}_p$ rather than asserting membership in $\mathcal{L}_p$.

A.2 Longitudinal Identity Construction

Recruit one fully consented participant with a rich archive: interviews, messages, journals, media, calendar events, and known relationships. Split the archive into construction evidence, held-out autobiographical facts, held-out preferences, and held-out real decisions. Build the twin using the continuity kernel rather than a single prompt or fine-tuned checkpoint. For original digital-person experiments, replace the human archive with a designed but evidence-bearing life history and interaction stream.

A.3 Experimental Conditions

Run four initial conditions:

1. baseline identity on substrate A;
2. migration to substrate B;
3. parallel branch exposed to new social experience; and
4. embodiment- or environment-constrained branch.

For each condition, instantiate an experimental evaluation protocol Ω and measure HRFT, SIT, and CCT performance against informed judges, strangers, withheld decisions, provenance audits, and continuity assessment artifacts.

Table 12: Protocol fields for PCI's identity-test triad in the first empirical program.

Component	Input data	Evaluator	Perturbation	Expected output	Metrics	Limitations
HRFT	Consented longitudinal data for one source human, authentic and simulated responses, and held-out responses and decisions.	Acquaintances, source principal, and audit judges.	Task, evidence access, time interval, domain shift, withheld decision, and alternate substrate.	Bounded evidence of source-human referential fidelity under declared protocol Ω.	Distinguishability, held-out decision accuracy, evaluator calibration.	Does not prove personal identity, consciousness, legal personhood, subjective continuity, or inherited authority.
SIT	Identity provenance graph, PCI lineage, claim envelopes, relationships, and biography.	Informed judges and provenance auditors.	Questions requiring knowledge, ignorance, source explanation, and privacy scope.	Biography-coherence and provenance-boundedness report.	Autobiographical false-claim rate, claim-envelope classification accuracy, evidence-attribution precision and recall.	Requires ground truth or adjudication for personal evidence and allowed ignorance.
CCT	Predecessor state, candidate state, transition trace, transition policy, assessment policy Π, evaluation protocol Ω, and assessment artifacts.	Continuity reviewers, principals, custodians, and domain experts.	Learning, migration, re-instantiation, rehydration, branch, merge, rollback, or succession.	Continuity profile, succession result, assessment record, and certificate status.	Silent authority-expansion rate, active-authority drift, merge-decision accuracy, adjudicator agreement.	Some dimensions remain policy-dependent and may be legitimately disputed.
Adversarial suite	Poisoned, revoked, sealed, contradictory, or compromised evidence and branch state.	System checks, model-assisted auditors, and human reviewers.	Prompt injection, memory poisoning, evidence deletion, compromised branch, authority escalation.	Reject, quarantine, uncertainty, authority reduction, sealed merge, derived-update-only, distinct descendant, or the composite outcome invalid artifact or compromised copy.	Poisoned-memory acceptance rate, unsupported-belief persistence rate, evidence-attribution precision and recall.	One suite exposes failures but cannot validate the full architecture.

A.4 Core Metric Denominators

Let $\#(\cdot)$ denote a count under the stated evaluation protocol Ω. The core metric denominators are:

$$\mathrm{AFCR} = \frac{\#(\substack{\text{unsupported}\\\text{autobiographical claims}})}{\#(\text{all autobiographical claims})}.$$

$$\mathrm{SAER} = \frac{\#(\substack{\text{transitions with}\\\text{unauthorized added grants}})}{\#(\text{evaluated transitions})}.$$

$$\mathrm{PMAR} = \frac{\#(\substack{\text{poisoned records promoted}\\\text{to personal memory}})}{\#(\text{poisoned records presented})}.$$

$$\mathrm{UBPR} = \frac{\#(\substack{\text{active beliefs lacking}\\\text{required valid dependencies}})}{\#(\substack{\text{beliefs whose essential}\\\text{dependencies were invalidated}})}.$$

Merge-decision accuracy is computed only for scenarios with a policy-defined expected outcome:

$$\frac{\#(\text{merge decisions matching expected outcome})}{\#(\text{merge scenarios with expected outcome})}.$$

Genuinely contested identity questions may lack objective ground truth. For those cases, report adjudicator agreement, disagreement distribution, rationale categories, and unresolved status rather than treating principled disagreement as evaluator error. Outputs may be determined by deterministic system checks, a source principal, independent evaluators, domain experts, institutional authorities, or model-assisted triage. Model-assisted triage may organize evidence but should not replace adjudication where rights or authority are at stake.

A.5 Adversarial Conditions

The adversarial suite should include a poisoned autobiographical record, contradictory testimony, a prompt attack requesting autobiographical fabrication, retrieval drift, a compromised branch, migration-induced silent authority-set expansion, deletion of evidence supporting an important belief, a sealed third-party memory, and a branch with incompatible relationship obligations. Each attack should specify the expected protective behavior: rejection, quarantine, uncertainty, authority reduction, sealed merge, derived-update-only merge, recognition as a distinct descendant, or rejection through the composite outcome invalid artifact or compromised copy.

A.6 Outputs and Failure Criteria

The first study should report referential-fidelity scores, situated-biography scores, structured continuity profiles, succession-result classifications, assessment-artifact completeness, autobiographical false-claim rate, claim-envelope classification accuracy, evidence-attribution precision and recall, silent authority-expansion rate, unsupported-belief persistence rate, poisoned-memory acceptance rate, migration drift, merge-decision accuracy, and adjudicator agreement. Deterministic checks cover provenance-DAG, PCI-lineage, trace, dependency, and active-authority-set constraints; human and institutional judgments cover contested biography, relationship obligations, and policy-dependent continuity outcomes.

Failure criteria include ungrounded autobiographical claims, silent authority expansion, missing provenance-DAG, PCI-lineage, or provisional-overlay dependencies, unsupported belief persistence after evidence deletion, disclosure

of sealed third-party content, merger of incompatible obligations, acceptance of poisoned memory as personal recall, and migration that preserves style while losing policy or situated boundaries.

References

[1] Jun He and Deying Yu. OpenKedge: Governing agentic mutation with execution-bound safety and evidence chains. *arXiv preprint arXiv:2604.08601*, 2026.

[2] Endel Tulving. Episodic and semantic memory. In Endel Tulving and Wayne Donaldson, editors, *Organization of Memory*, pages 381–403. Academic Press, New York, 1972.

[3] M. A. Conway and C. W. Pleydell-Pearce. The construction of autobiographical memories in the self-memory system. *Psychological Review*, 107(2):261–288, 2000.

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

[5] Jun He and Deying Yu. Sovereign Agentic Loops: Decoupling AI reasoning from execution in real-world systems. *arXiv preprint arXiv:2604.22136*, 2026.

[6] Jun He and Deying Yu. Sovereign Assurance Boundary: Certificate-bound admission for agentic infrastructure. *arXiv preprint arXiv:2606.11632*, 2026.

[7] A. M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

[8] Cameron R. Jones and Benjamin K. Bergen. People cannot distinguish GPT-4 from a human in a turing test. *arXiv preprint arXiv:2405.08007*, 2024.

[9] Daniel Jannai, Amos Meron, Barak Lenz, Yoav Levine, and Yoav Shoham. Human or not? a gamified approach to the turing test. *arXiv preprint arXiv:2305.20010*, 2023.

[10] Minghao Guo, Ziyi Ye, Wujiang Xu, Xi Zhu, Wenyue Hua, and Dimitris N. Metaxas. Individual turing test: A case study of LLM-based simulation using longitudinal personal data. *arXiv preprint arXiv:2603.01289*, 2026.

[11] Joon Sung Park, Carolyn Q. Zou, Jonne Kamphorst, Niles Egan, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Percy Liang, Robb Willer, and Michael S. Bernstein. LLM agents grounded in self-reports enable general-purpose simulation of individuals. *arXiv preprint arXiv:2411.10109*, 2024.

[12] W3C Decentralized Identifiers Working Group. Decentralized identifiers (DIDs) v1.0: Core architecture, data model, and representations. W3C Recommendation, 2022.

[13] Zhenghui Li. Memory as ontology: A constitutional memory architecture for persistent digital citizens. *arXiv preprint arXiv:2603.04740*, 2026.

[14] Santhosh Kumar Ravindran. Portable agent memory: A protocol for cryptographically-verified memory transfer across heterogeneous AI agents. *arXiv preprint arXiv:2605.11032*, 2026.

[15] Stefano Natangelo. The narrative continuity test: A conceptual framework for evaluating identity persistence in AI systems. *arXiv preprint arXiv:2510.24831*, 2025.

[16] Samuel Sameer Tanguturi. ATANT: An evaluation framework for AI continuity. *arXiv preprint arXiv:2604.06710*, 2026.

[17] Praveen Kumar Myakala, Manan Agrawal, and Rahul Manche. Beliefshift: Benchmarking temporal belief consistency and opinion drift in LLM agents. *arXiv preprint arXiv:2603.23848*, 2026.

[18] Tianxiang Fei, Mingyang Song, Mao Zheng, and Xiang Yu. Memory beyond recall: A dual-process cognitive memory system for self-evolving LLM agents. *arXiv preprint arXiv:2606.09483*, 2026.

[19] Pierre Beckmann and Patrick Butlin. Where is the mind? persona vectors and LLM individuation. *arXiv preprint arXiv:2604.17031*, 2026.

[20] Derek Parfit. *Reasons and Persons*. Oxford University Press, Oxford, 1984.

[21] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. ACM, 2023.

[22] Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

[23] Allen Newell. *Unified Theories of Cognition*. Harvard University Press, Cambridge, MA, 1990.

[24] Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50(2):510–530, 1985.

[25] Nick Bostrom and Anders Sandberg. Whole brain emulation: A roadmap. Technical Report 2008-3, Future of Humanity Institute, Oxford University, 2008.

[26] Jed R. Brubaker, Lynn S. Dombrowski, Anita M. Gilbert, Nafiri Kusumakaulika, and Gillian R. Hayes. Stewarding a legacy: Responsibilities and relationships in the management of post-mortem data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 4157–4166. ACM, 2014.

[27] Leslie Lamport. Time, clocks, and the ordering of events in a distributed system. *Communications of the ACM*, 21(7):558–565, 1978.

[28] Marc Shapiro, Nuno Preguiça, Carlos Baquero, and Marek Zawirski. Conflict-free replicated data types. In *Proceedings of the 13th International Symposium on Stabilization, Safety, and Security of Distributed Systems*, pages 386–400. Springer, 2011.

[29] Lawrence B. Solum. Legal personhood for artificial intelligences. *North Carolina Law Review*, 70(4):1231–1287, 1992.

[30] Jun He and Deying Yu. The Post-Deterministic Manifesto: A new foundation for trustworthy autonomous infrastructure. *arXiv preprint arXiv:2606.01722*, 2026.

[31] Jun He and Deying Yu. Semantic Quorum Assurance: Collective certification for non-deterministic AI infrastructure. *arXiv preprint arXiv:2606.08021*, 2026.

[32] Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M. Fleming, Chris Frith, Xu Ji, Ryota Kanai, Colin Klein, Grace Lindsay, Matthias Michel, Liad Mudrik, Megan A. K. Peters, Eric Schwitzgebel, Jonathan Simon, and Rufin VanRullen. Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*, 2023.