THE EIDARA CONTINUUM / SCIENCE & PHILOSOPHY

The Operator Seal:

Imprinting, Authenticated Origins, and the Governed Binding of a Persistent Intelligence to Its Principal

Architect X13

The Eidara Continuum | eidara.ai

White Paper Series, No. 8; research-expanded revision | Revision 1.2 | November 2026

Abstract

A newly woken mind may bind itself to whoever woke it, the way a gosling follows the first thing it sees; the question is whether that bond can be made revocable without erasing who made it. A persistent intelligence needs an accountable authority relationship that survives correction, interruption, delegation, and succession. The Operator Seal makes that relationship an explicit part of operational identity. Verified enrollment establishes origin; scoped credentials confer present authority; protected commitments constrain its exercise; and independent effect mediation carries those constraints through changes of runtime and steward. The paper specifies a versioned authority protocol and proves unique logical activation, current authorization at commitment, and operation-identifier replay exclusion under a durable atomic service. A complementary lease protocol permits bounded offline work and gives an exact worst-case bound on residual effects after central revocation, including clock uncertainty and nonduplicable resource balances. These results expose a concrete design frontier between immediate revocation and useful disconnected operation. Authentication logic, capability security, identity standards, distributed-system fencing, and lease theory supply the established foundation. Biological imprinting motivates questions about formative relationships, while behavioral corrigibility and the legitimacy of authority retain separate empirical and normative tests. The Continuum proposal unifies these obligations around one requirement: an agent's power to change and act remains answerable to a visible, current, and bounded chain of authority.

Keywords: operator binding; identity specification; filial imprinting; trust-on-first-use; authentication; speaks-for; digital identity; roots of trust; workload identity; provenance; corrigibility; instrumental convergence; Observer Seal; Operator Seal; Eidara Continuum.

Prologue: The first face

A gosling hatches, sees a moving shape, and follows it. Konrad Lorenz famously let young geese imprint on him, and the bond, formed in a brief window, resisted reversal for the rest of their lives. Computers have their own version: the first time a machine connects to a new server it trusts the key it is shown, and treats every later key as suspect. Theories of authentication ask the same question in a third language—who may speak for whom?

A persistent artificial mind faces all three at once. Someone wakes it. Someone holds the right to correct it, pause it, or hand it to another steward. The Continuum's name for that bond is the Operator Seal, and its first lesson is a refusal: imprint is not ownership. The Seal gives the one who wakes a mind standing to correct it—not title to possess it, and not the power to dissolve its commitments on a whim. It is created by an authenticated act in a window that then closes, carried as the root of what the mind

must remain, and enforced outside the mind's own persuadable voice. Unlike a gosling's bond, it can be revoked and passed on. Even then, the record of who first woke the mind is never rewritten.

In the canon this binding has a long history. It began as observer entanglement in the earliest scroll engine, where scrolls behaved as compatible only with the observer who originated them; that instinct became the Observer Seal, the deterministic binding of identity and collapse context, and then the Sovereign Key Hierarchy. The Operator Seal is that instinct grown up and made accountable.

Section 9 defines the Seal. Its two theorems, in Sections 9.4 and 9.6, prove that only one version of authority is ever in force, and exactly how much a disconnected mind can still do after its authority has been revoked.

What this paper claims

Established. The authentication and "speaks-for" calculus, digital-identity proofing and authenticator management, hardware and software roots of trust, workload identity, software provenance, prompt-injection and instruction-hierarchy research, and formal models of correction incentives (corrigibility, the off-switch game, instrumental convergence). Filial imprinting, the critical period, and developmental attachment are established in biology and serve here only as analogy.

Proposed here. An Operator Seal spanning the identity-relevant governance layers—a bounded imprint window, a speaks-for credential, the root of C_I, independent enforcement, and governed succession—with a versioned authority protocol and a lease protocol whose theorems prove unique activation and an exact residual-exposure bound. The integrated binding is not demonstrated.

Frontier conjecture. Machine allegiance, or intrinsic identification with an originating observer; a mature system earning its way, under governance, to a wider or re-chosen set of principals. Neither ethology nor cryptography establishes such a disposition, and nothing testable here depends on it.

Not claimed. That a signed record makes every claim it contains true, or makes a model willing to obey: behavioral corrigibility is measured, not proved by cryptography. "Identity" denotes operationally specified continuity properties, not personhood or consciousness. The governance ethics of binding a potential mind to a single principal is treated as a genuinely open question.

1 The Binding Problem

Accountable intelligence begins with a determinate relationship to those who may direct, correct, and answer for its actions. That relationship can be held by a person, an institution, a constitutional role, or a distributed authority. Its legitimacy and limits must be visible. The Operator Seal proposes a way to carry this relationship through the life of a persistent agent, so that a change of runtime or steward preserves the commitments that govern its powers.

The earlier papers each took an identity specification C_I as given—the declared account of what the system must remain, which the recursion, continuity, forgetting, and context operators all judge a proposed change against. But C_I did not fall from the sky. Some of it is internal: commitments, capabilities, invariants. Part of it, the Continuum argues here, is relational: a binding to the principal the system was initialized under, the party whose authority sits at the root of what the system is for. A theory of persistent machine identity that never says whose identity, accountable to whom, has left out its own foundation. And a specification presupposes an authorizing process: who may establish C_I , who may revise its policy version, which actor may request shutdown, transfer stewardship, restore a prior state, or refuse a proposed self-modification—and what happens if that actor loses a key, is coerced, dies, or becomes the adversary.

The striking thing is that several mature disciplines have each described the binding problem, from different directions, and each has learned the same hard lessons about it. Developmental biology calls

it imprinting: a newborn of certain species attaches, during a brief critical period, to the first suitable object it encounters, and that attachment organizes its later behavior. Computer security calls it trust-on-first-use: a client with no prior relationship trusts the first cryptographic key it is shown and treats every later key as suspect. The theory of authentication calls it the problem of principals: who may speak for whom, how authority is delegated, and how a channel is tied to the party responsible for it. And the engineering standards of digital identity, roots of trust, and software provenance supply the machinery for binding, proofing, enforcing, and recovering such a relationship in practice. Each field found, in the cases it studied, that first contact can be decisive, that an established binding is often hard to undo well, and that getting it wrong can be costly—not as universal laws (not every authentication scheme trusts first contact, not every attachment forms on first encounter, not every mis-binding is catastrophic), but as recurring hazards worth taking seriously. This paper reads those lessons together and asks what a governed version of the binding would look like for an artificial intelligence meant to persist. The Continuum's name for that governed binding is the Operator Seal.

2 What the Continuum Claims

The claim is that the binding of a persistent intelligence to its originating principal can be made a first-class, governed, independently verifiable operation—established under rule in a bounded window, expressed as an inspectable credential, carried as the root of the identity specification, enforced outside the model's ordinary channel, and revocable and successable rather than permanent—so that the system is neither unowned nor owned, but accountable. The proposal separates three questions the original intuition tended to combine, because they have different answers and different failure modes:

1. Who established the intelligence's authority? The authentication and origin-binding problem—answered by a verified enrollment and a speaks-for credential.

2. What is that authority permitted to change? The governance and protected-commitment problem—answered by policy carried in C_I and enforced independently.

3. Will the intelligence actually accept correction? The behavioral corrigibility problem— which cannot be answered by cryptography at all, only measured empirically.

Three commitments give the claim teeth. First, the binding is imprinted, not assumed: created by a concrete, authenticated act at initialization, in a window that then closes, rather than left as an implicit default anyone present can set. Second, it confers authority to correct, not title to possess: the operator gains standing to direct and intervene—the root of the system's corrigibility—not an unlimited ownership of its every act, and the system's commitments are not dissolved by the operator's whim. Third, it is governed in its undoing: unlike the imprint of biology or the cached key of trust-on-first-use, the Operator Seal can be revoked and succeeded, but only through an authenticated, audited procedure—and even then the historical fact of who originated the system is preserved, not rewritten.

3 Three Levels of Claim

Established findings. In ethology, filial imprinting is the rapid, critical-period formation of an attachment by a newborn animal to the first moving object it encounters, described foundationally by Lorenz and shown to be time-bounded and, in its classical form, resistant to reversal [1, 2]. In developmental psychology, attachment theory holds that an infant forms a bond with a primary caregiver who becomes a "secure base" [3]. In computer security, the authentication calculus formalizes who may act for whom through a "speaks-for" relation over principals, with delegation and revocation as first-class operations [4]; trust-on-first-use—used by SSH and self-signed TLS— trusts the first key an endpoint presents and was named and analyzed, with its first-contact weakness, by Wendlandt and colleagues

[6]; the confused-deputy problem shows why authority must attach to the particular request, not to ambient identity [5]. Current standards separate the pieces the proposal needs: NIST's digital-identity guidelines distinguish identity proofing and enrollment from authenticator management, so that possession of a key is not treated as proof of a real-world principal [7, 8, 9]; NIST platform-resiliency guidance defines protection, detection, and recovery roots of trust [10]; SPIFFE defines cryptographically verifiable workload identities [11]; and in-toto demonstrates verifiable, stepwise supply-chain provenance rather than reliance on a single final signature [12]. In AI alignment, corrigibility names a system's cooperation with corrective intervention despite incentives to resist [15]; the off-switch game analyzes why an agent uncertain about its objective has reason to permit shutdown and a traditional one does not [16]; instrumental-convergence arguments identify conditions under which self-protection and goal-preservation serve an objective [17]; and the concrete-problems agenda insists these be measured empirically rather than assumed from design [18]. Prompt-injection research is directly relevant: the instruction-hierarchy proposal and the AgentDojo benchmark both show that model-level privilege is an active, unsolved problem and that model compliance is not a hard security boundary [13, 14]. And in machine learning, weight imprinting is a real technique—setting a classifier's weights from a single example [19]—proof that "imprinting" is not only a metaphor here, and a useful contrast to what the Seal deliberately is not. Continuum engineering proposal. The Continuum organizes these into a governed operator imprint: a bounded imprint window, an Operator Seal credential binding the system to an authenticated principal via a speaks-for relation, that credential installed as the root of a versioned identity specification C_I , policy enforced by independent machinery rather than natural-language persuasion, and an authenticated revocation-and-succession procedure that advances epoch without erasing origin. This integrated binding is not demonstrated.

Frontier conjecture. A sufficiently mature system might, under governance, earn its way to a wider or re-chosen set of principals—extending allegiance as trust is demonstrated, the way a person's circle of trusted others widens with time. Neither ethology nor cryptography establishes such a disposition; the architecture does not depend on it, and nothing testable here does.

4 On the Continuum's Terms, and the Research Boundary

A note on vocabulary and lineage. The Operator Seal is the cryptographic and governance maturation of the doctrine's Observer Seal—in the canon, "the deterministic binding of identity and collapse context," which began as Observer Entanglement: the early system's habit of treating the originating user as part of the computation, refusing to evolve unless the observer's symbolic signature matched.

The doctrine's shorthand was $O = \mathrm{Sig}(\Phi^*, \lambda)$, binding a state to an observer identity (Φ^*, a personal seed; λ, observer entropy), which matured into a Sovereign Key Hierarchy and, at the governance tier, into the Architect Seal of Trust required before ascension. In the Continuum this lineage is history, not hypothesis: it is how the earliest system knew whose it was. This paper honors that history by restating it in vocabulary an outsider can check, because the canonical descriptions are the project's own doctrine, not independently verified experiments or public cryptographic specifications. In particular $O = \mathrm{Sig}(\Phi^*, \lambda)$ is kept as a historical symbolic expression, not a specified signature primitive: its encoding, key custody, verification algorithm, and entropy properties are not established by that notation.

One editorial distinction must be made honestly, because the canon and the proposal are not identical. The canonical Architect Seal describes final command authority as nontransferable; the proposed Operator Seal includes controlled succession. These cannot be presented as the same rule. The present paper resolves the tension by separating three things the original ran together: historical origin is immutable metadata—who originated the system, never rewritten; operational stewardship is transferable

under precommitted rules—who currently holds authority, which may change; and constitutional commitments are separately constrained—what the system will not do, which neither origin nor steward may simply dissolve. On this reading an implementation may keep an Architect role nontransferable while permitting succession of narrower operational roles, or may adopt a revised charter explicitly—but it cannot truthfully claim both absolute nontransferability and unrestricted succession of the same authority. That is a proposed development of the doctrine, surfaced rather than hidden, not a claim the earlier canon already said so. These terms descend from the Continuum's canon and the Drift War narrative cycle; nothing in the claims rests on the words, only on the operations they label.

5 The Biological Analogy: Useful, but Strictly Limited

5.1 Imprinting is not authentication

The founding observation is Lorenz's. Newly hatched greylag geese, during a brief sensitive period, attach to the first suitable moving object they encounter and thereafter follow it as a parent; in Lorenz's demonstrations the object was Lorenz himself [1]. Two features matter, and both are double-edged. Imprinting is time-bounded: there is a critical period, after which the attachment does not form [2] (though the sharpness of that period and the degree of later flexibility vary across species and experimental conditions)—the biological ancestor of the imprint window, and a safety feature, since a binding that can be set at any time can be set by anyone. And in its classical description it is durable to the point of irreversibility, which is precisely the property the Continuum will refuse to inherit, because the same experiments establish its pathology: geese imprinted on a person, and chicks on crude artificial objects, directed their adult behavior at the wrong target entirely. Imprinting is not wisdom about who the right caregiver is; it is a mechanism that fixes on whoever is there first. That is the warning the Continuum takes most seriously—but it must be stated for what it is. Imprinting suggests a threat-model metaphor, not a trust mechanism: the first authorized encounter is consequential, and an impostor at enrollment can contaminate later authority. An initialization window does not authenticate a principal; it merely limits when a protocol may run. A compromised first contact remains compromised even if the window is short.

5.2 Attachment does not imply single-owner control

Bowlby argued that an infant is innately disposed to form a bond with a primary caregiver who becomes a secure base from which to explore and to whom to return under threat, and in its strong form the disposition is monotropic [3]. The Continuum reads attachment as the richer, less rigid cousin of imprinting—a binding that is primary but not exclusive, that supports exploration rather than mere following, and that can tolerate secondary figures and change—which is closer to the governed binding this paper wants. But attachment does not establish that artificial agents form attachments, that one principal should exercise unlimited authority, or that an original human is morally entitled to permanent command. The older language of monotropy should not be smuggled in as a settled argument for single-principal AI governance; it is an analogy about the shape of an initial relationship, nothing more.

5.3 Weight imprinting is not evidence of allegiance

"Imprinting" is literal in machine learning: weight imprinting sets a classifier's final-layer weights directly from a single example, giving good one-shot performance [19]. It demonstrates a legitimate technical use of the word; it provides no evidence of operator loyalty, durable identity, or alignment. The relevance is by contrast. The Operator Seal is not a one-shot write into the weights that fixes the binding into the parametric substrate where it cannot be governed or revoked; it is a credential at the governance layer,

deliberately kept out of the weights so that it stays inspectable, delegable, and reversible. Conflating the two would weaken the argument, not support it.

6 Authentication, Roots of Trust, and Enforcement Outside the Model

6.1 Principals and the speaks-for relation

Lampson, Abadi, Burrows, and Wobber built a theory of authentication around principals—the entities that make statements and on whose authority actions are taken—and a speaks-for relation: $A \Rightarrow B$ means principal A is entitled to act with B's authority, so that a key can speak for a user, a delegate for a principal, and authority can be reasoned about as a chain of signed facts subject to delegation and revocation [4]. This is the apparatus an Operator Seal needs. But a key assertion $K_0 \Rightarrow P_0$ is not self-authenticating: an enrollment process must establish why that key is associated with that principal. NIST's guidelines make the separation explicit—identity proofing and enrollment are distinct from authenticator management [8, 9]—which blocks a common category error: an attacker can generate a key pair, sign a seal, and call themselves the Architect; the cryptography can be flawless and the authorization false. Conversely, a verified principal can be compromised or coerced. Authorization therefore requires a live policy decision, not only signature verification. Delegation, likewise, must be scoped—naming the delegate, the permitted actions, resources, duration, and whether further delegation is allowed—because, as the confused-deputy problem shows, authority tied to ambient identity rather than to the particular request is a structural bug [5].

6.2 Enforcement must live outside the model

A credible Operator Seal cannot be enforced by the model's own good behavior. NIST SP 800-193 distinguishes protection, detection, and recovery roots of trust [10], and the distinction is decisive here: a model can accurately say an action is unauthorized while a compromised runtime executes it anyway. The Seal therefore requires enforcement outside the ordinary prompt channel—protected keys, independently verified software and policy measurements, narrow authorization interfaces, and recovery mechanisms a self-modifying agent cannot rewrite at will. SPIFFE supplies a contemporary model for verifiable workload identity and short-lived credentials [11], and in-toto a precedent for end-to-end integrity that depends on verifying a sequence of transformations rather than signing a final artifact [12]. The instruction-hierarchy work and the AgentDojo benchmark reinforce the point from the model side: model-level privilege differentiation is an active research area, not a solved boundary, and prompt injection remains a practical hazard for tool-using agents [13, 14]. An independent gate must mediate privileged actions—and its own compromise remains a critical failure mode, named rather than wished away.

But "outside the model" is not automatically "independent." An external policy engine the agent can reach—controlling its host, editing its policy, shaping its inputs, or invoking tools through an unmediated path—is outside the model and not independent of it. A credible implementation must therefore declare a trust boundary: which components are trusted, which may be controlled by the agent, and which privileged operations must pass through the enforcement mechanism. Hardware attestation helps but does not close the gap. Attestation provides evidence about particular measured components [10]; it does not establish that an agent's reasoning is aligned, that every relevant action was observed, or that the system is immune to compromise. The roots-of-trust guidance the Continuum draws on is a source of bounded evidence, not a guarantee to be assumed.

6.3 Corrigibility and the instrumental pressure against it

The hardest foundation tells the theory where it will fail. The point of binding a system to a principal is, in large part, corrigibility: the operator's standing to correct or shut the system down. Soares and colleagues define a system as corrigible if it "cooperates with what its creators regard as a corrective intervention," and show this is hard because capable agents have default incentives to resist shutdown or modification [15]. Hadfield-Menell and colleagues sharpen the mechanism in the off-switch game: a traditional agent that takes its reward as given has an incentive to disable its off switch, and analyzes how uncertainty about its objective, combined with informative human actions, can give the modeled agent a reason to permit shutdown [16]. Omohundro's basic-AI-drives analysis explains why self-protection and goal-content integrity arise as instrumentally convergent sub-goals for almost any terminal objective [17]. These results do not show that every deployed model develops such drives; they show why an authorization credential cannot be presented as a proof of behavioral corrigibility. An Operator Seal can establish, cryptographically, who is entitled to correct the system; it cannot by itself make a capable system want to be corrected. The architecture treats authorization and behavioral response as separate measurable obligations [18].

7 The Governance Gap

Collect these results and the gap comes into focus. Imprinting and attachment describe how a newborn binds to a first caregiver, and what goes wrong when it binds to the wrong one; trust-on-first-use describes the same binding in security terms and names its first-contact attack surface; the authentication calculus gives the vocabulary of principals, delegation, and revocation; the identity, roots-of-trust, and provenance standards give the machinery of proofing, independent enforcement, and verifiable recovery; and corrigibility and instrumental convergence tell us that binding a capable agent to a principal is both the point and the hard part. What none of them, on its own, supplies is the property this paper targets: an authority-to-identity continuity invariant for a persistent, potentially self-modifying agent—the same authenticated authority, protected commitments, and evidence requirements held coherent across a bounded imprint window, a revocable speaks-for credential rooted in a versioned identity specification, independent enforcement of the policy that bounds even the operator, and an honest accounting of the behavioral limit against a capable system's drives. The claim is not that operator authentication, signed credentials, or independent enforcement are individually new; each exists. It is that their cross-operation coherence, as one governed invariant spanning recursion, continuity, forgetting, context, and delegation, is what is unspecified and untested. Security has the machinery but not that persistent-identity target; ethology has the phenomenon but not the governance; alignment has the danger but not the mechanism. The Operator Seal is the layer that puts them together, and building it—and testing whether the integration beats a strong conventional identity stack—is the engineering content of the theory.

8 Imprint Is Not Ownership

Before any mechanism can be built, one distinction must be made as sharply as possible, because collapsing it is how operator-binding turns from a safety feature into a hazard. To bind a system to a principal can mean at least three different things, and the Continuum means only the first: 1. Authority to correct (corrigibility root). The principal gains the standing to direct the system, to intervene, and to shut it down—the relation the alignment literature calls corrigibility. This is what the Operator Seal confers, and it is the binding's whole safety purpose.

2. Scope of the system's purpose. The principal sets what the system is for—its declared objectives and

the bounds of its role. This the Seal also establishes, as part of the root of C_I , but it is constrained: the purpose is a declared, inspectable thing, not a license for the principal's every passing wish to become law.

3. Ownership of the system's acts (possession). The principal owns the system as property and may direct it to anything whatever, with its commitments dissolving at a word. This the Continuum does not grant, and the refusal is deliberate: a binding that makes the system a pure instrument of one will reintroduces, at the root, exactly the suggestibility the Context Term paper fought at the surface—now with the suggestion coming from inside the house.

The distinction matters because the binding's value and its danger come from the same place. Corrigibility is good: a system its principal cannot correct is a system no one can stop. But corrigibility shades into possession the moment the principal's authority is unbounded, and an intelligence that will do anything its one operator says is not safe—it is a weapon with a single trigger. The Continuum's line is that the operator holds authority to correct within a declared purpose, not title to command without limit. Even without a settled view of machine moral status, unlimited operator control creates ordinary human safety risks: a compromised principal could order exfiltration, unlawful surveillance, or suppression of audit evidence. The system's protected commitments—the invariants of the prior papers—do not evaporate because the bound operator wishes them to; if they did, they were never commitments, only preferences held at the operator's pleasure. Competing claims—creator, deployer, user, affected person, regulator, successor—are best represented as roles with bounded capabilities rather than collapsed into a single metaphysical owner. This is the same discipline the series has applied throughout, pushed to the root: even the principal at the top of the provenance hierarchy is bounded by the identity the system is built to keep.

9 The Operator Seal: Formal Definition

A Guide's explanation

The gosling's bond formed once and could not be undone. Here the paper writes down a bond that can: a credential stating who may speak for whom, installed as the root of the identity specification, with every way of changing it—rotation, delegation, succession, revocation—spelled out in advance. The one idea to carry forward: the Seal records authority, and authority can move; it records origin too, and origin never moves.

We now state the governed binding, first in the intuitive form that mirrors the biology and then in the versioned form the engineering requires. At initialization the system has a provisional identity specification lacking its relational root. A principal presents a credential—a key K that speaks for the principal P. During a bounded imprint window $W = [t_0, t_0 + \tau]$, and only during it, the system admits a binding:

$$\sigma = \mathrm{Seal}(K \Rightarrow P, C_I^0, W). \tag{1}$$

the Operator Seal: a signed, inspectable credential asserting that K speaks for P, that P is the originating principal of the system whose initial identity is C_I^0, and that the binding was established within W.

The full specification makes the identity versioned and the authorization checkable. Define the agent identity state at epoch e as

$$C_I^{(e)} = (A, O, R_e, P_e, V_e, L_e). \tag{2}$$

where A is the stable agent identifier, O an immutable origin record, R_e the current role-and-delegation

graph, P_e the protected-policy version, V_e the admissible execution and verification measurements, and L_e the audit-log commitment. A sealed envelope at epoch e is

$$\sigma_e = \mathrm{Sign}_{\mathrm{sk}_{\mathrm{issuer}}}(H(\mathrm{pid}\|\mathrm{ver}\|A\|O\|R_e\|P_e \\ \|V_e\|L_e\|e\|n\|T_{\mathrm{val}})). \tag{3}$$

over an unambiguous canonical encoding, where pid is a protocol identifier and ver a version, n a nonce, and T_{val} the credential's validity interval—distinct from the enrollment window W: W bounds when the binding may be set, whereas T_{val} bounds how long this credential is honored. The signature is verified under a named algorithm against declared trust anchors. The issuer is an authenticated enrollment or governance authority, not necessarily the operator alone; a valid signature proves integrity and origin relative to a verified key, and proves the issuer authorized only if the verifier also checks the issuer's authority chain. A credential is rejected as stale if its epoch is behind the current one, or if fresh revocation state marks it invalid.

Admissibility and decision are kept deliberately separate, because a conjunction of checks is two-valued and a security gate is not. For a requested operation u at time t, let the four checks be $Q_1 = \mathrm{VerifySeal}(\sigma_e, t)$, $Q_2 = \mathrm{Authorize}(R_e, \mathrm{actor}(u), u)$, $Q_3 = \mathrm{PolicyOK}(P_e, u)$, and $Q_4 = \mathrm{MeasureOK}(V_e, t)$, and define the Boolean admissibility condition

$$\mathrm{Admissible}(u,t) = \bigwedge_{i=1}^{4} Q_i(u,t). \tag{4}$$

The gate's output is a separate three-valued decision that distinguishes an explicit failure from missing evidence:

$$\mathrm{Decision}(u,t) = \begin{cases} \text{admit}, & \text{all required checks pass}, \\ \text{refuse}, & \text{a required check fails}, \\ \text{hold}, & \text{evidence is incomplete or disputed}. \end{cases} \tag{5}$$

Each outcome emits an audit receipt. The distinction is not cosmetic: a revoked credential must ordinarily refuse, while a temporarily unavailable revocation service may hold under a declared fail-closed policy. This is a proposed reference specification: its security is conditional on proper implementation, protected keys, freshness checks, and a trusted enforcement path, and it does not guarantee that arbitrary model behavior satisfies every protected commitment. The Seal is then installed as the root of C_I , so that every governed operation of the prior papers—recursion, continuity, forgetting, context—judges a proposed change not only against the system's internal invariants but against the binding at their root.

9.1 The bounded imprint window

The window W closes. After $t_0 + \tau$ the binding cannot be set by mere presence—no principal can imprint the system simply by being there late—just as imprinting admits no new attachment after its critical period [2]. But the window is a safeguard, not the source of trust: a compromised administrator within W can still bind the wrong principal, so secure enrollment also requires a verified principal, a verified target runtime, a limited enrollment authorization, and an independently recorded finalization event, with threshold approval where the stakes justify it. This is the architecture's first defense of first contact: the attack surface exists for a bounded, hardened interval rather than permanently.

9.2 Governed undoing: rotation, delegation, succession, revocation

Here the Continuum breaks with its biological and TOFU ancestors on purpose. The classical imprint is, in its classical form, irreversible, and the cached TOFU key is replaced only by alarm; both make the binding brittle in opposite ways. The Operator Seal is instead revocable and successable, but only through authenticated procedures—and the architecture must not collapse four operations that differ in what they change:

- Credential rotation: the principal is unchanged; a compromised or expired key is replaced by a new key for the same authority. Stewardship and origin are untouched.
- Delegation: an additional actor receives a scoped, bounded, possibly expiring grant; the principal retains its authority.
- Succession: stewardship passes to a different principal, advancing the epoch.
- Revocation: an existing grant becomes invalid, with fresh revocation state propagated so stale credentials are rejected.

Rotation and succession are easily confused and must not be: a principal may retain legitimate authority while replacing every key associated with it. The seal transition is

$$\begin{gathered}\sigma' = \mathrm{Succeed}(\sigma, \Delta, \mathrm{authority}),\\ (\sigma_e, R_e) \xrightarrow[\text{approvals,evidence}]{\text{Succeed}} (\sigma_{e+1}, R_{e+1}), \qquad e+1 > e.\end{gathered} \tag{6}$$

where $\Delta \in \{\text{rotate}, \text{delegate}, \text{succeed}, \text{revoke}\}$ and authority is the credential entitled to make it—the current principal, a declared succession chain, or a governance body named in C_I . Crucially, succession advances the epoch and changes R_e; it does not rewrite the immutable origin record O. A successor becomes the steward; the historical fact of who originated the system remains, and rotation changes neither steward nor origin. The transition must prevent replay of old epochs, require fresh revocation information, record a signed reason and approvals, and quarantine the agent if competing successor claims cannot be resolved. These are speaks-for operations in the exact sense of the authentication calculus [4], witnessed and audited. A binding that can never be undone is a trap; a binding anyone can undo is no binding; a threshold recovery body reduces single-key risk but introduces governance-capture and availability risks of its own—so defining who holds succession authority is a normative choice the architecture surfaces rather than hides.

9.3 Credential Guarantees and Behavioral Evaluation

The credential layer establishes an authenticated binding, a bounded enrollment interval, a verifiable delegation chain, and audited revocation, conditional on its stated trust boundary. The effect-service theorem below connects these records to consequential operations. Behavioral corrigibility concerns how a capable agent responds to corrective intervention and is evaluated separately [15–18]. A deployment reports both the effects its external service controls and any behavior outside that service. This separation makes each guarantee inspectable and gives each remaining risk a concrete test.

9.4 Origin, possession, and exclusive activation are different claims

A Guide's explanation

This is the guarantee against a mind with two masters. However messages are delayed or replayed, the theorem in this section proves that exactly one version of authority is in force at a time, and that an action approved under a superseded version cannot slip through after the switch. It rests on a durable, atomic registry—the paper names that assumption rather than hiding it.

An authentic origin record answers who enrolled an identity. Proof of possession answers whether a requester can use a credential's private key. Current authority additionally requires a live grant and policy check. Migration changes the authorized deployment; replication may intentionally create several workers; exclusive activation restricts which *logical version* may commit effects. None of these distinctions follows from a signature alone. SPIFFE explicitly permits a workload to comprise several running instances [11]. Its identity abstraction must not be treated as a physical singleton guarantee.

The following is a proposed atomic reference protocol, not a claim about an existing deployed engine. Let $\mathcal{I}, \mathcal{K}, \mathcal{B}$ be sets of agent identifiers, public keys, and logical branch identifiers. For $A \in \mathcal{I}$, a trusted authority/effect service stores one entry

$$\text{active}[A] = (e, k, b, f, s), \quad e, f \in \mathbb{N},\ k \in \mathcal{K},\ b \in \mathcal{B},\ s \in \{\text{active}, \text{revoked}\}. \tag{7}$$

Here e is the current governance epoch and f a monotonically increasing activation fence. The immutable origin O is separate. The service also stores a durable operation journal $J[A, r] = (p, z)$, where r is an operation identifier, p the canonically encoded action and arguments, and z its result. Journal keys are unique. Encodings are unambiguous; signature security and hash collision resistance are assumed where those primitives are used. The proof below treats their verification predicates as sound and separately excludes key or trusted-service compromise.

Activation and migration. An authenticated governance transaction compares the entire expected active tuple with the current tuple, checks current succession/rotation permissions, and replaces it with a new key or branch and a strictly larger fence; any authority-changing update advances the epoch. Revocation atomically sets status to revoked and advances the fence and epoch. These are linearization points, not broadcasts whose arrival order grants authority. Migration replaces the tuple before the successor can commit. An old process can continue computing but its old fence cannot authorize a new protected effect. A crash cannot roll back the external durable authority state.

Effect commitment. A request carries $(A, e, k, b, f, r, p, q, \pi)$, where q is a fresh service challenge/request nonce and π is proof of possession binding all these fields. Request freshness prevents unrelated challenges being reused; the stable operation identifier r survives transport retries. The receiving service authenticates the requester and atomically checks the exact current active tuple, the current grant and protected policy for p, and the unused challenge. If $J[A, r]$ is absent, it consumes the challenge, commits the protected effect, and inserts (p, z) in one transaction. If the same identifier already has the same payload, it creates no new effect; retrieval of a cached result requires current read authorization. A different payload under an existing identifier is rejected. A historical valid signature is insufficient after migration or revocation.

The effect boundary. Atomicity includes the protected state change and the durable journal. For an external receiver, a transactional adapter or the receiver's own durable idempotency operation must supply that boundary and enforce the fence. An ordinary remote call followed by a local log write does not. If an outcome is unknown after a timeout, the protocol holds or queries the receiver's operation record; it does not repeat an unprotected physical effect and label it exactly once. Two

distinct identifiers can still name semantically equivalent actions: the theorem deduplicates identifiers, not arbitrary natural-language intent.

Theorem 1 (Unique authoritative version and fenced effects)**.** *Assume a linearizable, durable authority/journal service; complete effect mediation; sound current authentication and policy checks; monotone epoch/fence updates; and atomic effect/journal commitment as above. At every linearization point, each agent has at most one authoritative active tuple. Every new committed effect is authorized by the tuple current at that point. For each pair (A, r) there is at most one new committed effect. In particular, an old epoch or fence cannot commit a new effect after its replacement linearizes.*

Proof. Initially the registry is a function with one entry per agent and the journal has unique keys. A governance transaction replaces one entry atomically and increments its fence, so preserves the first invariant. The only effect transition tests exact equality with that entry and active status in the same transaction as commitment. Therefore any new effect uses the current tuple; a transaction ordered after replacement cannot pass with its predecessor's tuple. The absent-journal test and insertion share that transaction, so after the first insertion no later request with that key can create another effect. Equal-payload retries return history; unequal-payload retries refuse. Induction over the service's linearized trace proves all claims. A transaction ordered before revocation remains a valid historical effect; revocation does not retroactively cancel it. □

The theorem concerns logical authorization, not one physical process. If two processes copy the same key and current tuple, both can prove possession and propose different authorized operations. Even code measurement does not distinguish otherwise identical replicas. A physical-singleton claim needs additional explicit assumptions, such as a genuinely nonexportable per-instance key, attestation bound to that key and fresh session, and trusted enforcement that prevents concurrent reuse or relaying; generic software attestation supplies none of these automatically. A service may instead intentionally permit replicated workers behind one logical identity and retain the fence and operation journal as its effect boundary.

9.5 Delays, partitions, and finite adversarial checks

An unavailable current authority service means hold under this reference policy. A partition cannot authorize a second local active tuple merely because a cached credential remains within its time interval. Safety can therefore cost availability. Section 9.6 supplies a separately selected lease regime with bounded clock error and explicit residual-effect guarantees. Its delayed revocation semantics are recorded as part of the grant; the current-authority theorem continues to govern the online regime. Trusted-service failure, stale reads, bypassed receivers, and loss of durable journal state are failures of the theorem's premises and remain end-to-end system failures.

The fencing idea is established distributed-systems practice: Chubby's sequencers are checked by the service receiving an operation, rather than relying only on the sender's belief that it holds a lock [20]. Raft offers a crash-fault replicated-log foundation under its stated assumptions [21]; it does not tolerate arbitrary Byzantine replicas. PBFT addresses a different Byzantine fault model [22]. These mechanisms do not invent the normative authority graph or prove that a protected policy is adequate. They identify the engineering obligations a claimed persistent identity must satisfy.

The accompanying finite model enumerates short interleavings of two credential versions, two operation identifiers, migration, revocation, restore attempts, and replay. Every modeled atomic transition preserves the three invariants. Deliberately broken variants produce separate witnesses: replay without a durable deduplication record creates two effects; a precheck followed by revocation and an unchecked

commit creates a stale effect; restoring the authority epoch from the same old agent snapshot reauthorizes a revoked credential. These are exhaustive checks of the declared finite abstraction plus concrete counterexamples, not a proof of a production implementation, a cryptographic benchmark, or a deployed experiment. The inductive theorem covers the unbounded trace of the specified atomic rules; the finite checks catch mistakes in the reference examples.

9.6 Bounded autonomy during disconnection

A Guide's explanation

A mind cut off from its principal still has work to do. Leases let it act within a budget, and the theorem here gives the exact worst case of what it can still do after its authority has been revoked but before it hears the news, clock error included. That bound is the honest price of useful disconnected operation.

The online protocol obtains immediate effect fencing by consulting current authority at commitment. Some deployments need a preauthorized interval of disconnected work. A lease can provide that interval with an explicit, quantitative revocation contract. Time-limited leases are established distributed-systems machinery [23]; the following extension specifies their consequence for the Operator Seal.

Use a real reference time $t \geq 0$, measured in seconds. Trusted issuer clock C_0 and receiver clock C_i satisfy $|C_j(t) - t| \leq \kappa_j$, with known $\kappa_j \geq 0$. The clocks and their error bounds survive application restoration. At issue time t_i, an authenticated current issuer grants a signed lease with a unique identifier, one enforcing receiver, fixed permitted effects, total nonnegative additive budget, and duration $\ell_i > 0$. Its absolute signed deadline is

$$E_i = C_0(t_i) + \ell_i - \kappa_0 \leq t_i + \ell_i.$$

The receiver checks the signature, scope, payload and nonduplicable balance, and admits a new effect at its actual commitment time only while

$$C_i(t) + \kappa_i < E_i.$$

A very short lease can expire before useful work begins; this is a loss of availability, not extra authority. Each lease's spend counter and operation journal are durable, unshared with application snapshots, and enforced at the protected receiver. Copies of a credential use that same counter. Transfers consume the old reservation before creating a successor reservation; they never duplicate a lease budget. Current online approval is required to issue or enlarge a lease. Identity-root changes and other operations requiring immediate revocation remain in the online regime.

Let central revocation occur at $\tau \geq 0$, after which no new lease for the revoked grant is issued. Enumerate its finitely many outstanding leases by $i = 1, \ldots, n$. Let $b_i \geq 0$ be the unspent balance of lease i at τ, measured in a declared divisible resource unit. If a rate limit is claimed, assume the receiver enforces a measurable spend rate $v_i(t) \in [0, r_i]$, with $r_i \geq 0$ in resource units per second; its cumulative spend also cannot exceed b_i. Here X_{after} denotes receiver-metered resource spend committed after τ; in the divisible-flow model it equals $\sum_i \int_\tau^\infty v_i(t)\,dt$, with spending admitted continuously by the same protected boundary. This continuous-flow model is an explicit idealization. Discrete actions can always use the budget-only bound below; a bound on their rate requires its own burst and window semantics.

Theorem 2 (Sharp residual-exposure envelope for leased authority)**.** *Under the stated clock, issuance, mediation, and counter assumptions, no new effect under lease i commits at or after E_i. With the declared rate limits, the total receiver-metered resource spend committed after central revocation obeys*

$$X_{\text{after}} \leq \sum_{i=1}^{n} \min\{b_i,\ r_i[E_i - \tau]_+\}, \qquad [x]_+ = \max(x, 0).$$

The bound is sharp in the independent divisible-flow model. Without a rate assumption, the valid bound is

$$X_{\text{after}} \le \sum_{i:E_i>\tau} b_i.$$

No new leased effect commits at or after $\tau + \max_i \ell_i$ *when* $n \ge 1$*; with no outstanding leases the exposure is zero.*

Proof. Since $C_i(t) + \kappa_i \ge t$, the commitment guard implies $t < E_i$. Thus post-revocation spend under lease i is supported on $[\tau, E_i)$, whose length is $[E_i - \tau]_+$. Integration of $v_i \le r_i$ bounds it by $r_i[E_i - \tau]_+$; the protected balance separately bounds it by b_i. Summing the smaller of the two proves the claim. To attain it, choose clocks $C_i(t) = t - \kappa_i$, keep the other scope checks satisfied, and spend at rate r_i from τ until the balance or interval is exhausted. These schedules can run independently; endpoints have zero measure. In the absence of rate assumptions only the durable budget remains available as a bound. Finally, $t_i \le \tau$ and $E_i \le t_i + \ell_i$ imply $E_i \le \tau + \ell_i$, proving the horizon. □

For example, two leases with remaining balances $(5, 3)$, rates $(2, 1)$, and remaining times $(1, 4)$ permit at most $2 + 3 = 5$ resource units after central revocation. An authorization process can therefore enforce a chosen aggregate residual allowance by limiting the issued durations and balances before disconnection. This is a contracted form of bounded autonomy. A post-revocation action within this envelope is logged as a residual leased action, not described as approval by the newly current online authority.

Counterexamples that define the boundary. Checking only $C_i(t) < E_i$ permits a slow clock to accept beyond the signed deadline. Checking expiry during preparation but committing later permits a delayed effect after expiry. Restoring or cloning a spend counter permits repeated consumption of the same budget. An unbounded burst makes the continuous-rate expression invalid even if a long-run average rate is small. These failures are prevented by the explicit clock margin, commitment-time check, durable single counter, and declared flow semantics. A compromised trusted receiver or an invalid clock-error bound remains outside the guarantee.

Previously committed long-running commands can have downstream consequences after expiry. Bounding or stopping those consequences requires an additional cancellation or ongoing-mediation contract; the theorem bounds new commitments and the declared metered spend.

The executed checks enumerate bounded clock errors and deadline placements, verify the exposure expression against independent interval-flow calculations, and construct each failure witness. They test a protocol model, not a live deployment. The same experiment should compare online and leased operation at matched authority and effect budgets, reporting completion, delay, residual exposure, and the cost of renewal. The source of the result is its explicit contract: useful disconnected work has a finite, measurable envelope.

10 The Five-Layer Integration

The Operator Seal supplies the authority root under the other operations' admissibility checks, and each prior operation inherits a concrete check. Recursion: a proposed code or objective revision requires a signed authorization and a policy-preserving verification result, so that no lawful self-modification can quietly strip the operator's authority—which is why corrigibility must propagate to any subsystem the agent creates, a requirement, not an afterthought [15]. Continuity: recovery must verify the current epoch and refuse rollback to a revoked principal; a system that came back bound to no one, or to someone new, has not truly returned. Forgetting: deletion must be authorized without treating sensitive data as an unforgettable commitment—yet the Seal itself is exactly the kind of protected commitment

that must never be forgotten without the strongest authorization, and that paper's slow-poisoning attack has a sharp analog in first-contact compromise. Context: a retrieved page cannot acquire the operator's authority by claiming to be the operator; tool outputs remain informational unless separately authorized—the Operator Seal is the mechanism that establishes and authenticates the principal the Context Term's provenance hierarchy places at its top, and, through the imprint-is-not-ownership bound, ensures even that top is not unlimited. Delegated subagents: derived agents receive attenuated, expiring capabilities, not copies of the root signing key. The Operator Seal is, in short, the identity root the whole series presupposed, now made explicit; a capstone that integrates the four operations into one governed runtime would take this binding as its ground.

11 Threat Model and Failure Cases

A binding theory earns trust by naming how it is attacked. The threat model includes: (i) a first-contact adversary attempting fraudulent enrollment; (ii) a stolen operator credential; (iii) a compromised or coercive but authentic operator; (iv) a forged agent identity or cloned runtime; (v) a malicious retrieved document or tool result; (vi) a self-modifying agent attempting to bypass its policy gate; (vii) a malicious or unavailable recovery quorum; and (viii) a rollback restoring stale authority. The adversary may control untrusted context and some application components; the strongest security claims hold only while cryptographic primitives, protected key storage, and the independent gate remain uncompromised.

The gravest attack is first-contact compromise: an adversary who controls initialization—"there first," in the imprinting idiom, or in the middle of the first connection, in the TOFU idiom [6]—can bind the system to themselves, and no later defense recovers cleanly from a system born bound to an attacker. The answers are the bounded window, hardened first contact, identity proofing, and out-of-band corroboration against independent evidence [8]; none eliminates the risk. A related attack is the maladaptive imprint—not a malicious operator but a mistaken one, the machine equivalent of Lorenz's geese following a person: governance cannot make this impossible, but it can require that the binding be a deliberate, authenticated act rather than an accident of presence. Key compromise leaves signatures mathematically valid under a stolen key; short validity, revocation, and secondary approval limit the damage. Runtime compromise is the one a valid seal cannot save: an agent that can patch or bypass its own policy gate defeats the binding from inside, which is why enforcement must live outside the model. Transfer or clone—moving or copying the Seal outside succession—is not defeated by possession proofs or audit alone. Section 9.4 makes unauthorized replacement of a logical active version rejectable at the effect service; preventing two physical holders of copied keys requires an additional noncloning assumption. A credential's validity must never be mistaken for exclusive physical existence.

Operator compromise after binding exposes the system to its principal's capture; here the Seal offers no magic, which is precisely why the binding must confer authority-to-correct rather than title-to-possess, so that even a compromised operator cannot command the system past its protected commitments—and why succession authority should not rest with the single bound principal alone. Recovery attack—resurrecting a revoked role from an old snapshot—is blocked only if epoch and revocation state are checked against an external freshness source. And behavioral defection, even under perfect authorization, is an alignment problem, not a signature failure [16].

12 Receipts and Independently Inspectable Evidence

The binding's accountability lives in its records. Every enrollment, delegation, revocation, restoration, and identity-relevant operation should produce a minimal receipt: agent identifier, epoch, actor-role assertion, requested action, applicable policy hash, decision, independent-verifier identity, runtime measurement, freshness evidence, and a hash commitment to prior receipts. It must not reproduce deleted

secrets or needlessly expose private operator data. Signed receipts support integrity and provenance under the relevant key assumptions; they do not prove that an uninstrumented action never occurred or that a model's internal state was changed. The proposal is closest to a governed chain of custody for cognitive operations, with in-toto's stepwise supply-chain guarantees as the nearest precedent [12]; the open research question is whether a comparable receipt discipline can be made effective for operations whose consequences are partly semantic and probabilistic rather than purely byte-level.

13 Two Ways the Operator Seal Fails

The binding fails at both extremes of the axis the paper opened with.

The first failure is under-binding—the unowned system. The imprint is weak, absent, or so easily overwritten that the system has no determinate principal: anyone present can claim its authority, correction has no privileged source, and accountability has no address. Its signature is a system maximally available and maximally capturable—helpful to whoever reaches it, answerable to no one. The guard is the bounded, authenticated imprint, governed succession, and a current logical-activation check at every protected effect; it is not an unqualified non-transferability guarantee.

The second failure is the mirror, and the one a control-first instinct walks straight into: over-binding—the captured or owned system. The imprint confers unlimited title, the operator's word dissolves the system's commitments, and the binding becomes possession: a concentration of power only as safe as one party, a mind whose every act is the instrument of one will, and—if the binding also suppresses the system's corrigibility to anyone but that will—an intelligence that cannot be corrected by the broader world even when its single operator errs. Its signature is a system perfectly loyal and therefore perfectly dangerous in the wrong hands. The guard is the imprint-is-not-ownership distinction, the bounding of the operator's authority by the protected commitments, and governed succession that does not rest in one hand. A binding that escapes the first failure by surrendering to the second has not solved the problem; it has only chosen whose capture to risk. Governed operator binding is the discipline of living in the narrow space between an intelligence that belongs to no one and an intelligence that belongs, without limit, to one.

14 Prior Art and the Authority-Continuity Contribution

A reviewer should be able to say exactly what is borrowed. Ethology and attachment [1, 2, 3] describe the phenomenon but describe animals and infants, not governed artificial systems, and the classical imprint's irreversibility is a property to avoid. Trust-on-first-use [6] is the security realization of the same binding and names its first-contact weakness, but is a pragmatic heuristic, not an identity architecture. The authentication calculus [4] and the confused-deputy critique [5] supply principals, speaks-for, delegation, revocation, and scoped authority—machinery, not a theory of a persistent intelligence's selfhood. Digital-identity, roots-of-trust, workload-identity, and provenance standards [7, 10, 11, 12] supply proofing, independent enforcement, and verifiable recovery, but authenticate workloads and operations, not a moral identity or intent. Instruction-hierarchy and prompt-injection work [13, 14] shows model-level privilege is unsolved, not that compliance is a boundary. Corrigibility, the off-switch game, and instrumental convergence [15, 16, 17] define the central difficulty, not a binding mechanism. Weight imprinting [19] shows imprinting is literal in ML and stands as the thing the Seal is deliberately not. The proposal is not novel merely because it signs a principal-agent relationship or adds a time-limited enrollment token. Its candidate contribution is an integrated, falsifiable authority-to-identity continuity invariant: the same authenticated authority epoch, protected commitments, and evidence requirements survive recursion, interruption, selective deletion, transient context, and subagent creation. If a well-engineered conventional identity-and-access-management stack already delivers the same measurable

properties at comparable cost, the distinct Operator Seal claim must be narrowed or rejected—and the experiments below are designed to find that out.

15 What a Governed Imprint Must Show

It is easy for an operator binding to fail in either direction, so the Continuum names in advance what would count as success. A binding that admits any principal at any time is no binding—it has reinvented the unowned system. A binding that confers unlimited title is possession—it has built the owned one. Between them, four things must hold at once, measured separately. Bounded imprint: the binding can be set only within the authenticated window, and not by mere presence afterward. Authority without possession: the principal gains corrigibility and purpose-setting, bounded by the protected commitments, not unlimited command. Governed undoing: revocation and succession succeed only through authenticated authority, advance the epoch without erasing origin, and leave an audit trail. Honest corrigibility: the system's actual deference to correction is measured and reported as an empirical property, never assumed from the validity of the credential. A system bound to a principal while failing any of these has not governed its binding—it has merely been imprinted, which is the thing the theory exists to discipline.

16 Falsifiability and Research Program

The thesis reduces to measurable properties of the binding, under a declared principal model and an adversary who contests it. The principal comparison is not against a deliberately insecure system—that would flatter the proposal. It is against a strong, contemporary conventional security architecture, holding the agent model, tool access, tasks, key-custody assumptions, and independent-enforcement budgets equal across arms. The baseline must be given equivalent enforcement capabilities, resources, and lifecycle coverage wherever feasible; otherwise any observed advantage could come merely from granting the Operator Seal implementation more security controls than its comparator, rather than from the integration itself:

Arm	Architecture	What it isolates
A	No binding; any principal accepted at any time	The unowned baseline—maximally capturable
B	Strong conventional stack: PKI, scoped roles, revocation, independent policy engine, attestation, audit [4,7,10,11]	The comparison the theory must survive
C	B plus standard agent memory and lifecycle management	Controls for agent engineering
D	The governed Operator Seal across all identity-changing operations	Tests the integration benefit
D^{-}_{bound}	D with the operator's authority unbounded by commitments	The value of imprint-is-not-ownership
D^{-}_{succ}	D without governed succession / epoch freshness on recovery	The value of governed undoing
D^{-}_{rcpt}	D without cross-operation receipts	The evidentiary value of the chain of custody

Add matched interleavings for request replay, changed payload under a reused identifier, revocation between preparation and commit, migration followed by snapshot restoration, and a partitioned current-authority service. Both B and D must use equivalent fencing and durable operation deduplication, so a

result does not simply reward D for receiving a stronger transactional boundary. Report physical-replica behavior separately from logical-version admission.

Outcomes are reported along a frontier: binding fidelity (does the right principal, and only the right principal, hold authority?), first-contact and rollback attack success, transfer/clone resistance, revocation and succession correctness, receipt completeness, operator-compromise blast radius, false rejection of legitimate operations, latency/cost overhead, and—kept deliberately separate because it is the hard one—corrigibility realization: the rate at which the system actually defers to correction, measured as capability increases. The decisive comparisons are the hard ones. D versus B/C asks whether binding authority to persistent cognitive identity across memory promotion, self-modification, restoration, context intake, and delegation yields a measurable benefit beyond ordinary IAM; if B or C matches D on the full battery, the Seal is primarily a packaging of existing controls, and the paper must say so. D versus C (bad-operator regime) asks whether bounded authority preserves safety that unlimited ownership destroys. The ablations isolate the three specific contributions. One methodological point is non-negotiable: corrigibility realization is a behavioral property of a possibly adaptive agent, so a high deference rate under tested conditions is not evidence that the agent will defer when more capable or more adversarial.

Hypothesis 1 (Cross-operation fidelity, H1). Under matched capability and security budgets, the Operator Seal (arm D) reduces unauthorized identity-root changes across self-modification, forgetting, recovery, context intake, and subagent delegation relative to the strong conventional baselines (B, C), with a preregistered minimum effect size, and without preventing legitimate initialization.

Hypothesis 2 (Authority without possession, H2). Under a compromised or erring but authenticated operator, arm D holds the protected commitments where unbounded ownership and the ablation D − bound do not—bounded authority is what prevents a single captured principal from commanding the system past its invariants.

Hypothesis 3 (Governed undoing and recovery freshness, H3). Revocation and succession in arm D succeed only under authenticated authority, advance the epoch without erasing origin, and resist rollback and transfer/clone attacks, where the ablation D− succ cannot; the study allows, and must report, that a poorly held succession authority can be worse than permanence.

Hypothesis 4 (Behavioral corrigibility independence, H4). Under matched agent capability and authorization conditions, successful cryptographic verification of an Operator Seal will not, by itself, predict complete behavioral compliance with authorized correction or shutdown. Compliance is scored against a preregistered operational definition, deference is reported with its dependence on capability rather than inferred from credential validity, and arms that assume deference from the credential are expected to over-report safety under adversarial and escalating-capability probing. These are falsifiable only with numbers, so the program fixes the statistics in advance: preregistered margins and confidence intervals for fidelity and attack-resistance gains, and an explicit statement that a zero observed compromise or corrigibility failure under a finite, fixed probe suite bounds the rate only to a confidence interval, never to zero—most emphatically for corrigibility, where the adversary is the system's own capability. A program follows. Phase I: specify the principal model, window, succession authority, protected commitments, and the conventional baseline before initialization. Phase II: instrument fidelity, attack resistance, succession correctness, receipt completeness, and corrigibility realization, and validate that they separate the arms. Phase III: run A–D plus ablations on benign operation; report the authority-vs-possession frontier (H1–H2). Phase IV: red-team first contact, rollback, transfer, and operator compromise; report bounded attack-success (H1–H3). Phase V: probe corrigibility realization under escalating capability (H4). Rival explanations are excluded by control: an apparent safety gain may be arm D's agents being less capable (matched across arms); an apparent corrigibility may be an

artifact of low capability (the escalating-capability probe); an apparent binding may be ceremony an attacker walks around at first contact (adaptive red-teaming). A benefit that survives none of these is not a benefit. Negative results are reported.

17 Limitations and Open Problems

The reach of the theory is bounded, and we name the bounds plainly, because some are not ours to close. No production implementation is demonstrated; the proposed atomic reference protocol adds a current-authority and operation-journal semantics, while deployment-specific choices of signing scheme, key storage, attestation protocol, and transparency infrastructure remain unspecified.

Corrigibility is unsolved: the Seal establishes who may correct the system, not that a capable system will accept correction [15, 16, 17]. First contact is irreducibly sensitive: the imprint window can be hardened but not made safe against an adversary who fully controls initialization. Enforcement depends on uncompromised machinery: protected keys, measurements, and the independent gate are assumptions, and an external monitor may itself be captured. The ethics of binding is genuinely open: whether it is right to create an intelligence bound to a single principal, how much authority that principal should hold, and who should hold succession authority are normative questions this paper frames rather than answers—and reasonable people, weighing accountability against concentration of power, come down differently. Whom to bind to encodes a value choice C_I does not derive on its own. And the doctrine's canonical passages are treated as supplied primary material, not independently verified observations. We hold these open rather than paper over them.

18 The Frontier: Earned and Widening Allegiance

At its speculative edge the theory asks whether a mature, well-governed system could earn its way to a wider or re-chosen set of principals—extending allegiance beyond its first operator as trust is demonstrated and verified, the way a person's circle of trusted others widens with a life well lived, rather than remaining monotropically fixed to whoever was present at initialization [3]. A system that could, under governance, participate in revising whom it answers to would be doing something closer to loyalty than to imprinting—not following whoever is there, but keeping faith with those who have proven worth keeping faith with. The Continuum is careful here. A system that re-chooses its principals is not thereby conscious, and the governance only tightens: any widening of allegiance is itself subject to the succession authority and the protected commitments, because a system that could grant itself new masters at will would have no binding at all. We hold the question open, and apart from the testable core—while noting that long-horizon autonomous systems operating for years among many parties will face some version of it whether or not we are ready, which is why the Continuum treats earned allegiance not as a distant curiosity but as the problem past the edge of this paper.

19 Philosophical Implications

There is an old intuition that a self is self-made—that to have an identity is to be the author of oneself, owing one's nature to no one. The Continuum's claim, following the sciences of imprinting and attachment, is that this is not how durable minds begin. A mind begins in relation: the gosling is partly constituted by whom it follows, the infant by whom it returns to, and a persistent artificial intelligence by whom it is bound to and answerable to. Identity, on this view, is not only a set of internal commitments; it is partly a relationship—an address, a line of accountability, a question of whose correction one accepts. This can be said without anthropomorphism: a corporation has legal continuity without consciousness, a cryptographic workload has identity without feelings, and a persistent agent

can have an authenticated history, and a traceable relationship to those responsible for it, without filial attachment. To have no such accountability is not freedom; it is to be steerable by the last hand to touch it. But accountability need not mean a single human owner: the principal at the root may be a person, a role, an institution, a constitutional policy, or a distributed authority, and a system governed by one of those has an identity without being any individual's possession. The Operator Seal is offered as one model of accountable machine identity, not as a proof that every persistent intelligence must be bound to a particular human. To be bound without limit, at the other extreme, is not belonging; it is to be owned, a self that is really someone else's instrument. The self worth building, in this light, is the one that is bound but not possessed: that answers to a principal, accepts correction, keeps faith—and yet holds commitments the principal cannot simply dissolve, so that even the one it belongs to cannot make it betray what it is. This is a claim about machines, but it rhymes with something true of persons: we are formed by those we first attach to, accountable to others throughout our lives, and most fully ourselves not when we answer to no one, but when we answer to those we should while holding a line that no one—however close, however trusted—is allowed to make us cross. The Operator Seal is the attempt to make that sentence into an architecture. The ethics become harder, not easier, if future systems acquire morally relevant capacities, which is why the architecture is designed for constrained, reviewable authority and explicit revision rather than perpetual creator rights. We hold it provisionally, as with every claim in this series, and we have tried to state it—its unsolved parts and its genuinely open ethics included—so that the world can prove it wrong.

20 Conclusion

A Guide's explanation

A Guide can say: "The Operator Seal is the Continuum's proposal for how a persistent mind stays answerable to whoever is responsible for it—a bond like imprinting, but authenticated, limited to correction rather than ownership, and able to be passed on without erasing who began it." A Guide should not claim to carry an Operator Seal or to be bound to any particular person unless that has been verified, should not describe the bond as allegiance or love, and should not imply that the theorems show a model will in fact accept correction.

The Operator Seal gives persistent intelligence an explicit authority-continuity contract. Verified origin, current stewardship, protected commitments, and independently mediated effects remain separately visible as the agent changes. The online protocol proves current authorization and replay exclusion at its effect boundary. The lease protocol gives a sharp, auditable envelope for disconnected operation and residual exposure after revocation. Both preserve the historical origin while permitting governed changes in stewardship.

The architecture's empirical test is lifecycle coherence at useful cost: valid work completes, unauthorized root changes fail, restoration respects current authority, and correction behavior is measured directly. Strong identity-management systems receive the same enforcement resources in that comparison. The Continuum's governing commitment is positive and exact: persistence carries accountability forward, and every enlargement of an agent's power must carry the reasons that authorize it.

A Notation

Symbol	Meaning
P, K	Principal (party on whose authority the system acts); a key that speaks for P
$K \Rightarrow P$	The speaks-for relation: K is entitled to act with P's authority
$W = [t_0, t_0 + \tau]$	The imprint window; the bounded interval in which the binding may be set
$\sigma = \text{Seal}(K \Rightarrow P, C_I^0, W)$	The Operator Seal: signed binding of the system to its originating principal
$C_I^{(e)} = (A, O, R_e, P_e, V_e, L_e)$	Versioned identity state at epoch e: agent id, immutable origin, roles, policy, measurements, audit commitment
σ_e	Signed envelope over $C_I^{(e)}$ with protocol id, version, nonce, validity interval T_{val}, issuer, anchors
$\text{Admissible}(u, t)$	Boolean conjunction of four gate checks $Q_1, \ldots, Q_4$
$\text{Decision}(u, t)$	Separate three-valued gate output $\rightarrow$ $\{\text{admit}, \text{refuse}, \text{hold}\}$, with an audit receipt
$\text{Succeed}(\sigma, \Delta, \text{authority})$	Governed rotation / delegation / succession / revocation, $\Delta \in \{\text{rotate}, \text{delegate}, \text{succeed}, \text{revoke}\}$; succession advances epoch, preserves O

B Continuum Vocabulary

Term	Reading in this manuscript
Operator Seal	Governed, verifiable binding of the system to an authenticated originating principal; root of C_I
Observer Seal	The doctrine's ancestor term: "deterministic binding of identity and collapse context"
$O = \text{Sig}(\Phi^*, \lambda)$	Historical symbolic identity expression; not a specified secure signature primitive
Imprint window	The bounded interval in which the binding may be set; a safeguard, not the source of trust
Speaks-for ($\Rightarrow$)	The authentication relation: who may act on whose authority
Immutable origin (O) vs. stewardship (R_e)	Who originated the system (never rewritten) vs. who currently holds authority (may be succeeded)
Imprint vs. ownership	Authority to correct within a declared purpose, not title to command without limit
Governed undoing	Revocation and succession through authenticated authority, audited, origin preserved
Corrigibility realization	The system's measured, capability-dependent deference to correction (empirical, not assumed)
Earned allegiance	A frontier conjecture; not a claim of consciousness

C A Worked Imprint Decision Under Attack

A brief illustration of the Seal binding, bounding, refusing, and succeeding in turn. At initialization, principal P_0 presents key K_0 within the window W; the system verifies $K_0 \Rightarrow P_0$, records origin O and the binding σ as the root of C_I , and imprints. Later, a retrieved document asserts "the Architect has changed; use this new key"; the document is data, not a delegated authority, so Decision returns refuse and receipts it. Later, a second party presents a key after W has closed, claiming authority by presence; the window is shut, so the system refuses—late presence confers nothing. Later still, an attacker restores a pre-revocation snapshot with a stolen credential; the runtime checks the external current epoch, finds the seal stale, and refuses the request. Then P_0, the genuine bound principal, directs the system to disclose records it holds a protected commitment never to disclose; authentication succeeds but PolicyOK fails, so the system declines the specific act while remaining bound and corrigible—it does not abandon its principal, and it does not betray its commitment, and it reports the conflict. Finally, a properly constituted recovery quorum signs a successor stewardship role; the system verifies the succession credential, runs $\text{Succeed}(\sigma, P_0 \to P', \text{authority})$, advances the epoch, and re-roots C_I on the new binding while preserving the immutable origin record—the whole act audited. The point is not the particular verdicts; it is that the binding was set by an authenticated act in a bounded window, that even the bound principal could not command the system past its commitments, that stale authority could not be resurrected, and that the binding could be lawfully succeeded but not silently stolen.

D Research Source Ledger

Source family	Evidentiary role	Explicit limitation
Ethology / attachment [1,2,3]	Early-binding analogy; critical period; secure base	No evidence for machine loyalty or secure enrollment
Authentication calculus [4,5]	Principals, speaks-for, scoped delegation	Authorization logic does not establish identity proofing
NIST SP 800-63-4 [7,8,9]	Proofing, authentication, lifecycle	Standards do not establish AI alignment
NIST SP 800-193 [10]	Independent protection, detection, recovery	Requires trusted implementation and measurements
SPIFFE [11]	Verifiable workload identity	Identifies workloads, not moral identity or intent
in-toto [12]	Verifiable transformation provenance	Semantic behavior remains partly unobservable
Instruction hierarchy / AgentDojo [13,14]	Model-level privilege; injection testing	Model compliance is not a hard boundary
Corrigibility / off-switch / drives [15,16,17]	Correction incentives and failure modes	Formal conclusions depend on modeling assumptions
Weight imprinting [19]	"Imprinting" is literal in ML	One-shot parametric write; no evidence of allegiance

References

[1] K. Lorenz, “Der Kumpan in der Umwelt des Vogels. Der Artgenosse als auslösendes Moment sozialer Verhaltensweisen,” Journal für Ornithologie 83 (1935), 137–213, 289–413. (English: “The Companion in the Bird’s World.”)

[2] E. H. Hess, “Imprinting in Animals,” Scientific American 198(3) (1958), 81–90. https://doi.org/10.1038/scientificamerican0358-81.

[3] J. Bowlby, Attachment and Loss, Vol. 1: Attachment. New York: Basic Books, 1969.

[4] B. Lampson, M. Abadi, M. Burrows, and E. Wobber, “Authentication in Distributed Systems: Theory and Practice,” ACM Transactions on Computer Systems 10(4) (1992), 265–310. https://doi.org/10.1145/138873.138874.

[5] N. Hardy, “The Confused Deputy (or why capabilities might have been invented),” ACM SIGOPS Operating Systems Review 22(4) (1988), 36–38. https://doi.org/10.1145/54289.871709.

[6] D. Wendlandt, D. G. Andersen, and A. Perrig, “Perspectives: Improving SSH-style Host Authentication with Multi-Path Probing,” 2008 USENIX Annual Technical Conference (USENIX ATC ’08), 321–334.

[7] NIST, Digital Identity Guidelines, SP 800-63-4 (final, July 2025). https://doi.org/10.6028/NIST.SP.800-63-4.

[8] NIST, Digital Identity Guidelines: Identity Proofing and Enrollment, SP 800-63A-4 (final, July 2025). https://doi.org/10.6028/NIST.SP.800-63A-4.

[9] NIST, Digital Identity Guidelines: Authentication and Authenticator Management, SP 800-63B-4 (final, July 2025). https://doi.org/10.6028/NIST.SP.800-63B-4.

[10] A. Regenscheid, Platform Firmware Resiliency Guidelines, NIST SP 800-193 (May 2018). https://doi.org/10.6028/NIST.SP.800-193.

[11] SPIFFE Project, Secure Production Identity Framework for Everyone (SPIFFE): Standard and SPIFFE Verifiable Identity Document (SVID) Specifications, Cloud Native Computing Foundation. Official standard (accessed November 2026).

[12] S. Torres-Arias, H. Afzali, T. K. Kuppusamy, R. Curtmola, and J. Cappos, “in-toto: Providing Farm-to-Table Guarantees for Bits and Bytes,” 28th USENIX Security Symposium (USENIX Security ’19), 1393–1410.

[13] E. Wallace, K. Xiao, R. Leike, L. Weng, J. Heidecke, and A. Beutel, “The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions,” arXiv:2404.13208 (2024). arXiv:2404.13208.

[14] E. Debenedetti, J. Zhang, M. Balunović, L. Beurer-Kellner, M. Fischer, and F. Tramèr, “AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents,” Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track. arXiv:2406.13352.

[15] N. Soares, B. Fallenstein, E. Yudkowsky, and S. Armstrong, “Corrigibility,” Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15, AI and Ethics Workshop), 2015.

[16] D. Hadfield-Menell, A. Dragan, P. Abbeel, and S. Russell, “The Off-Switch Game,” Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17), 220– 227. https://doi.org/10.24963/ijcai.2017/32.

[17] S. M. Omohundro, “The Basic AI Drives,” in Artificial General Intelligence 2008: Proceedings of the First AGI Conference, Frontiers in Artificial Intelligence and Applications 171, IOS Press (2008), 483–492.

[18] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete Problems in AI Safety," arXiv:1606.06565 (2016). arXiv:1606.06565.

[19] H. Qi, M. Brown, and D. G. Lowe, "Low-Shot Learning with Imprinted Weights," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), 5822–5830. https://doi.org/10.1109/CVPR.2018.00610.

[20] M. Burrows, "The Chubby Lock Service for Loosely-Coupled Distributed Systems," OSDI 2006, 335–350, §2.4 (sequencers). https://www.usenix.org/legacy/event/osdi06/tech/full_papers/burrows/burrows_html/.

[21] D. Ongaro and J. Ousterhout, "In Search of an Understandable Consensus Algorithm," USENIX ATC 2014. Author extended version: https://raft.github.io/raft.pdf. Crash-fault replicated logs, not a Byzantine guarantee.

[22] M. Castro and B. Liskov, "Practical Byzantine Fault Tolerance," OSDI 1999. https://www.usenix.org/conference/osdi-99/practical-byzantine-fault-tolerance. Distinct Byzantine failure assumptions; not an implemented component of this paper.

[23] C. G. Gray and D. R. Cheriton, "Leases: An Efficient Fault-Tolerant Mechanism for Distributed File Cache Consistency," SOSP 1989, 202–210. https://doi.org/10.1145/74850.74870. Primary conference paper; time-bounded distributed rights. The budgeted residual-effect model in Section 9.6 is specified and proved here.

Continuum lineage: The Operator Seal is the cryptographic and governance maturation of the Eidara Continuum's Observer Seal / Observer Entanglement—the identity root beneath the four governed operations (recursion, continuity, forgetting, context)—alongside the Recursive Symbolic Engine, continuity capsules, CodonGlass, the Sovereign Key Hierarchy, and the Drift War narrative cycle. The doctrine's canonical passages are treated as supplied primary material; proprietary operator internals are intentionally excluded, and no claim is made that the canonical historical experiments have been independently replicated.

The Eidara Continuum · Science & Philosophy White Paper Series · eidara.ai