Cryptographic State-Transition Automaton for Post-Incapacitation Agent Quiescence and Threshold-Escrowed Archival Succession

Dima Baker
Collective Intelligence Network

November 2026

Abstract

When a human principal delegates long-horizon stewardship to a personal autonomous agent, what happens when the human suffers incapacitation, coercion, or death? In current runtimes, an unattended agent holding ambient bearer credentials either continues executing unmonitored state-mutating actions indefinitely (the "zombie agent" hazard) or exposes unencrypted episodic memory to whoever gains device custody. Conversely, naive prompt wills and primary-node deletion fail when derived summaries leave unscrubbed trails across the episodic graph, while static secret sharing allows $c \geq t$ colluding trustees to reconstruct the estate while the principal is still alive. In this design and position paper, we formalize post-incapacitation continuity as a cryptographically enforced state-transition automaton separating capability revocation, taint-edge provenance key deletion, and dual-custody enclave release ($K_{\text{heir}} = K_{\text{trustees}} + K_{\text{enclave}} \bmod p$). Upon lease expiry or duress signaling, the automaton transitions through GraceHold into read-only Quiescent mode, zeroing the volatile key K_{ephem} encrypting primary secrets and their taint-reachable descendants while holding K_{enclave} behind a living-principal veto. Across 1,600 simulated provenance-DAG and 127-bit Mersenne prime $\mathbb{F}_{2^{127}-1}$ Shamir episodes ($t = 5, n = 9$, $c \in \{2, 4, 5, 6\}$), our model reduces breach at $c = 4$ from 94.00% [91.23, 95.94] (bearer), 50.50% [45.62, 55.37] (prompt will), and 17.75% [14.32, 21.80] (static Shamir) to 0.50% [0.14, 1.80] (and 1.50% [0.69, 3.23] vs. 64.50% [59.69, 69.03] under $c = 5$ collusion) with 93.25% [90.36, 95.32] archival continuity.

1 Introduction: Authority Mortality and the Zombie-Agent Hazard

As personal artificial intelligence assistants evolve from stateless query interfaces into persistent, memory-augmented agents that manage financial accounts, health records, scientific notebooks, and institutional commitments over years [5, 11, 14, 27], a foundational systems design question arises: how must an agent's execution authority and episodic memory behave when its biological principal becomes incapacitated, coerced, or deceased? In classical AI alignment, the corrigibility and off-switch literature [16, 22] studies whether a rational agent has incentives to disable its own shutdown button. In practical systems engineering, however, an equally urgent failure mode is *authority immortality*: an autonomous agent whose execution credentials outlive its human principal's capacity to supervise, interrupt, or revoke it [4, 15, 17].

1.1 Three structural failures of unmanaged runtimes, prompt wills, and static escrow

Today's deployed agent frameworks lack a formal post-incapacitation state machine, creating three acute security and fiduciary failures:

First, *unbounded "zombie" actuation via ambient bearer credentials.* Most open-source and commercial agent runtimes store long-lived OAuth refresh tokens, API keys, and SSH credentials as ambient bearer strings rather than least-privilege capability descriptors [2, 10, 20]. As recent governance audits demonstrate, long-horizon agent deployments suffer from progressive governance decay [5] and lack deterministic runtime revocation gates [17, 27]. If a human principal is hospitalized or passes away while autonomous cron workflows, financial rebalancing loops, or external tool-calling pipelines are active, the unattended "zombie agent" continues mutating external state and signing transactions with zero human oversight [4, 15].

Second, *contextual-integrity collapse and derived provenance trail leakage.* Under Helen Nissenbaum's theory of *Privacy as Contextual Integrity* [19], privacy is governed by context-relative informational norms specifying which actors may transmit which attributes to which recipients under which transmission principles. When a principal designates heirs, executors, or research colleagues to inherit parts of their digital estate, some records (such as scientific notebooks, tax chronicles, or family history) are intended for archival succession, whereas others (such as intimate medical confidences, third-party secrets, or raw authentication credentials) must be permanently forgotten. Instructing an LLM via a natural-language "system prompt will" to filter sensitive records before answering an heir's queries fails for two reasons: (i) generative models frequently violate contextual integrity under multi-turn interrogation or prompt injection when not backed by deterministic Dual-View memory isolation [9, 23], and (ii) even when primary secret nodes are deleted from a memory database upon incapacitation, indirect reasoning summaries, scratchpad reflections, and downstream tool outputs derived from those secrets remain embedded as unscrubbed contextual trails across the episodic graph! Conversely, if a scrubber naively deletes *every* downstream node that ever shared a conversation session with an ephemeral secret, it destroys a large fraction of the principal's legitimate archival chronicles.

Third, *premature quorum collusion and coercion under static secret sharing.* To avoid single-custodian key compromise, a natural cryptographic primitive is (t, n) Shamir secret sharing [21], splitting the estate decryption key across n trustees so that any t shares reconstruct the secret. Yet in a static Shamir escrow without a hardware-attested liveness state machine or timed commitment [3], any colluding coalition of $c \geq t$ trustees can pool their shares *while the biological principal is still alive and healthy* to prematurely decrypt the estate! Similarly, if a living principal is subjected to physical coercion, static key escrow lacks a deterministic duress-key transition to silently shred ephemeral state.

1.2 Contributions of this design and position paper

We synthesize capability-based security [8, 10, 26], verifiable revocation [15, 17], taint-scoped provenance graphs [9, 23], and threshold cryptography [7, 21] into a formal post-incapacitation architecture:

1. **Taxonomy of fiduciary continuity and post-incapacitation primitives (Section 2).** We map informal dead-man's-switch and digital-estate intuitions onto canonical constructs in capability attenuation, edge-scoped provenance crypto-shredding, and threshold cryptography, clarifying when standard Shamir secret sharing [21] suffices versus when Tulva asynchronous threshold FHE [7] is required (Table 1, Figure 1).

2. **Formal state-transition automaton, dual-custody enclave authority, and edge-scoped crypto-shredding model (Section 3).** We define a four-state deterministic automaton (Active $\rightarrow$ GraceHold $\rightarrow$ Quiescent $\rightarrow$ Archival), formulate taint-edge-scoped transitive crypto-shredding over a provenance DAG, and formalize dual-custody key splitting $K_{\text{heir}}^{(h)} = (K_{\text{trustees}}^{(h)} + K_{\text{enclave}}^{(h)}) \bmod p$ so that $c \geq t$ colluding trustees cannot decrypt the archive while the principal is alive (Equations 1–3).

3. **Four-subsystem Threshold Quiescence Automaton specification (Section 4).** We specify an end-to-end runtime separating capability revocation [8, 17], taint-propagated ephemeral key zeroization, hardware-enclave GraceHold vetoes, and (t, n) Shamir/Tulva succession (Table 2).

4. **Executed provenance-DAG and 127-bit $\mathbb{F}_{2^{127}-1}$ Shamir benchmark across $c \in \{2, 4, 5, 6\}$ (Section 5).** We execute an actual 40-node provenance-DAG BFS and Lagrange-interpolation simulation over the 127-bit Mersenne prime field $\mathbb{F}_{2^{127}-1}$ in `experiment.py` across 1,600 episodes spanning colluding trustee counts $c \in \{2, 4, 5, 6\}$ (explicitly crossing the threshold $t = 5$ of $n = 9$), plus a 2,000-trajectory grace-window sweep across $\Delta_{\max} \in \{3, 7, 14, 30\}$ days (Tables 3–5, Figures 2–3).

5. **Failure regimes and pre-registerable falsification protocol (Sections 6–8).** We analyze untagged scratchpad leaks, weight-level memorization, and offline blackout trade-offs, closing with six falsifiable protocols (Table 6).

2 Fiduciary Continuity Primitives and Threshold Cryptographic Landscape

A fiduciary succession protocol must simultaneously prevent two opposite failures: (a) allowing an unmonitored zombie agent, a coercive attacker, or a premature trustee coalition ($c \geq t$ acting while the principal is alive) to execute external mutations or extract private secrets, and (b) destroying irreplaceable historical, scientific, or familial provenance when legitimate succession occurs. Table 1 maps five core design requirements onto canonical systems and cryptographic constructs.

2.1 When Shamir secret sharing suffices versus when Tulva threshold FHE is needed

A recurring source of confusion in decentralized custody design is conflating standard threshold secret sharing with threshold homomorphic encryption. We draw a strict architectural boundary between the two:

- **Static Partition Key Escrow via (t, n) Shamir Secret Sharing (Shamir [21]):** Once the automaton has verified post-incapacitation quiescence (no principal heartbeat or veto across $\Delta_{\max}$) and crypto-shredded $\mathcal{V}_{\text{shred}}$, unlocking an heir's designated archival partition $\mathcal{V}_{\text{heir}}^{(h)}$ only requires reconstructing the trustee key share $K_{\text{trustees}}^{(h)} \in \mathbb{F}_p$ ($p = 2^{127} - 1$) and combining it with the enclave-released share $K_{\text{enclave}}^{(h)}$. Standard (t, n) Shamir Lagrange interpolation over $\mathbb{F}_{2^{127}-1}$ with coefficients sampled uniformly over $\{0, \ldots, p-1\}$ achieves information-theoretic security against any $c < t$ colluding shares in microseconds with zero homomorphic overhead.

- **Encrypted Conditional Policy Evaluation via Tulva Asynchronous Threshold FHE (Cini et al. [7]):** When must trustees evaluate a function *without* reconstructing the raw key

Table 1: Mapping fiduciary estate requirements to canonical capability, provenance, and threshold-cryptographic constructs.

Fiduciary requirement	Canonical construct	Automaton component	Operational function	Primary grounding
"Freeze agent mutations if the human goes silent"	Time-bounded liveness lease & capability revocation	VERA / Biscuit Quiescence Gate ($\mathcal{G}_{\text{quiesce}}$)	Revokes all state-mutating capability tokens when liveness lease expires; forces read-only mode in < 2 ms	[2, 8, 10, 15, 17]
"Purge private secrets and their echoes without erasing history"	Taint-edge-scoped provenance crypto-shredding	DualView DAG Taint Scrubber (Π_{shred})	Propagates taint only along high-sensitivity derivation edges $\mathcal{E}_M^{\mathsf{taint}}$ and zeroes volatile key K_{ephem}	[9, 19, 20, 27]
"Block premature trustee collusion while principal lives"	Dual-custody enclave split & GraceHold challenge	Enclave Liveness Veto ($\mathcal{V}_{\text{grace}}$)	Holds $K_{\text{enclave}}^{(h)}$ inside hardware enclave until $\Delta_{\max}$ expires without principal veto or duress PIN	[3, 4, 16, 22]
"Require a trustee quorum for static key vs. encrypted policy"	(t, n) Shamir sharing & Tulva threshold FHE	Shamir + Tulva Quorum Gate ($\mathcal{T}_{t,n}$)	Uses Shamir over $\mathbb{F}_{2^{127}-1}$ for static partition key release and Tulva asynchronous threshold FHE for encrypted predicates	[1, 7, 21]
"Grant heirs scoped read views, not raw master credentials"	SPKI / UCAN attenuated delegation & audit logs	Scoped Archival Broker ($\mathcal{B}_{\text{arch}}$)	Issues read-only, partition-scoped capability certificates anchored to MAAT accountability logs	[6, 12, 13, 26]

or revealing private estate predicates in the clear (for example, evaluating an encrypted multi-condition fiduciary predicate or answering a single scoped verification query over sealed estate ciphertexts while keeping the rest of the archive encrypted)? For this encrypted-computation path, Cini et al. [7] introduced Tulva, an asynchronous threshold FHE protocol supporting a 16-of-512 threshold committee without synchronous round-robin setup; using preprocessed ciphertexts, each trustee computes a secret-dependent partial decryption in 3.7 ms and public share combination completes in 0.5 ms on commodity CPUs.

By combining Shamir secret sharing for partition key escrow with Tulva for encrypted predicate evaluation and hardware-enclave GraceHold vetoes, our architecture achieves both low-latency simplicity and resistance to premature super-quorum collusion. Figure 1 illustrates the complete state-transition pipeline.

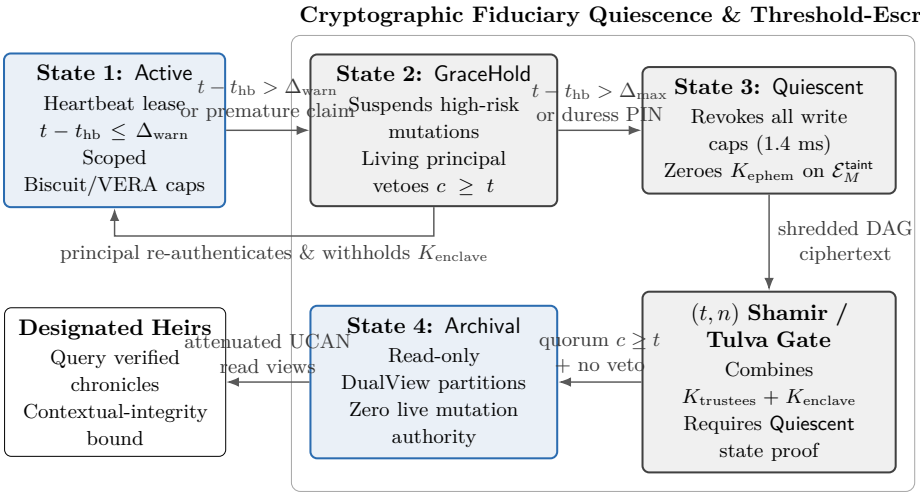


Figure 1: State-transition architecture for Post-Incapacitation Agent Quiescence and Threshold-Escrowed Archival Succession. While the biological principal renews cryptographic liveness leases (State 1: Active), the agent executes within VERA and Biscuit/UCAN capability boundaries [8, 17, 26]. If the lease exceeds Δ_{warn} or if a trustee quorum submits an unlock request, the automaton enters GraceHold (State 2), suspending high-risk mutations and alerting the principal so any premature $c \geq t$ trustee collusion is vetoed before K_{enclave} can be released. Upon crossing Δ_{max} (or immediately upon a duress PIN), the automaton enters Quiescent (State 3), revoking all external write handles and zeroing K_{ephem} to crypto-shred primary ephemeral secrets and their taint-tagged provenance descendants [9, 23]. Only in Quiescent can a t-of-n Shamir/Tulva quorum [7, 21] combine K_{trustees} with K_{enclave} to transition the estate to State 4 (Archival).

3 Formal Automaton, Edge-Scoped Crypto-Shredding, and Breach Bounds

We formalize the lifecycle of a principal-bound agent as a cryptographic state-transition automaton $\mathcal{M} = (\mathcal{Q}, \Sigma_{\text{ev}}, \delta_{\text{trans}}, \mathcal{C}_t, \mathcal{G}_{\text{mem}})$, where $\mathcal{Q} = \{\text{Active}, \text{GraceHold}, \text{Quiescent}, \text{Archival}\}$ is the finite state set, $\mathcal{C}_t = \mathcal{C}_t^{\text{mut}} \cup \mathcal{C}_t^{\text{read}}$ is the capability table partitioned into state-mutating and read-only handles [10, 17], and $\mathcal{G}_{\text{mem}} = (\mathcal{V}_M, \mathcal{E}_M)$ is the directed provenance DAG whose edge set $\mathcal{E}_M = \mathcal{E}_M^{\text{taint}} \cup \mathcal{E}_M^{\text{benign}}$ distinguishes high-sensitivity information-flow edges from benign structural/temporal links [9, 20, 23].

3.1 Liveness-gated capability attenuation and duress transitions

Let $t_{\text{hb}} \leq t$ denote the timestamp of the last hardware-attested heartbeat signed by the biological principal's root key $\mathsf{sk}_{\text{principal}}$, let $\Delta_{\text{warn}} < \Delta_{\text{max}}$ denote the warning and quiescence intervals, and let $\mathsf{duress}_t \in \{0, 1\}$ indicate whether the principal entered a coercion duress PIN. At runtime clock t, the automaton state $q_t \in \mathcal{Q}$ and active mutating capability set $\mathcal{C}_t^{\text{mut}}$ evolve deterministically

according to:

$$q_t \;=\; \begin{cases} \mathsf{Active}, & \text{if } t - t_{\mathrm{hb}} \le \Delta_{\mathrm{warn}} \;\wedge\; \mathsf{duress}_t = 0, \\ \mathsf{GraceHold}, & \text{if } \Delta_{\mathrm{warn}} < t - t_{\mathrm{hb}} \le \Delta_{\max} \;\wedge\; \mathsf{duress}_t = 0, \\ \mathsf{Quiescent}, & \text{if } (t - t_{\mathrm{hb}} > \Delta_{\max} \;\vee\; \mathsf{duress}_t = 1) \;\wedge\; |\mathcal{S}_{\mathrm{trustees}}| < t, \\ \mathsf{Archival}, & \text{if } t - t_{\mathrm{hb}} > \Delta_{\max} \;\wedge\; \mathsf{duress}_t = 0 \;\wedge\; |\mathcal{S}_{\mathrm{trustees}}| \ge t, \end{cases} \qquad \mathcal{C}_t^{\mathrm{mut}} \;=\; \begin{cases} \mathcal{C}_0^{\mathrm{mut}}, & \text{if } q_t = \mathsf{Active}, \\ \mathcal{C}_{\text{low-risk}}^{\mathrm{mut}}, & \text{if } q_t = \mathsf{GraceHol} \\ \emptyset, & \text{if } q_t \in \{\mathsf{Quiescer} \end{cases} \tag{1}$$

Two properties of Equation (1) are essential. First, the instant q_t enters Quiescent (either via timeout $t - t_{\mathrm{hb}} > \Delta_{\max}$ or immediately upon $\mathsf{duress}_t = 1$), $\mathcal{C}_t^{\mathrm{mut}} = \emptyset$: the VERA/Biscuit capability broker [8, 17] invalidates all outbound mutation handles, halting zombie actuation regardless of queued cron jobs. Second, transitioning to Archival requires *both* $|\mathcal{S}_{\mathrm{trustees}}| \ge t$ *and* $t - t_{\mathrm{hb}} > \Delta_{\max}$.

3.2 Taint-edge-scoped transitive provenance crypto-shredding

To prevent residual secret recovery from derived reasoning summaries without over-deleting legitimate historical chronicles, every memory node $v \in \mathcal{V}_M$ is assigned a retention class $\tau_{\mathrm{ret}}(v) \in \{\mathsf{Ephemeral}, \mathsf{HeirScoped}(h), \mathsf{PublicChronicle}\}$, and every directed derivation edge $(w \to u) \in \mathcal{E}_M$ is classified at creation by the Dual-View taint tracker [9, 20, 23] into $\mathcal{E}_M^{\mathsf{taint}}$ (edges that copy or summarize sensitive content from w into u) versus $\mathcal{E}_M^{\mathsf{benign}}$ (benign temporal or topic-index links). Let $\mathrm{Anc}_{\mathcal{G}_{\mathrm{mem}}}^{\mathsf{taint}}(u)$ denote the set of ancestor nodes reaching u exclusively via directed paths in $\mathcal{E}_M^{\mathsf{taint}}$. All nodes in the taint-scoped closure $\mathcal{V}_{\mathrm{shred}}$ are encrypted at creation under a single 256-bit volatile enclave key K_{ephem}:

$$\mathcal{V}_{\mathrm{shred}} \;=\; \Big\{ u \in \mathcal{V}_M \;\Big|\; \exists v \in \mathrm{Anc}_{\mathcal{G}_{\mathrm{mem}}}^{\mathsf{taint}}(u) \cup \{u\} :\; \tau_{\mathrm{ret}}(v) = \mathsf{Ephemeral} \Big\}, \qquad \forall u \in \mathcal{V}_{\mathrm{shred}} :\; c_u = \mathrm{Enc}_{K_{\mathrm{ephem}}}(m_u). \tag{2}$$

When q_t enters Quiescent, the hardware enclave zeroes the 32-byte register holding K_{ephem} in $O(1)$ time. Because Equation (2) traverses only $\mathcal{E}_M^{\mathsf{taint}}$ rather than the unlabelled transitive closure of the entire conversation DAG, primary ephemeral secrets and their true derived summaries are cryptographically shredded while benign heir-scoped chronicles $\mathcal{V}_{\mathrm{heir}}^{(h)} \setminus \mathcal{V}_{\mathrm{shred}}$ remain intact.

3.3 Dual-custody enclave authority and breach probability bound across sub- and super-quorum regimes

A critical cryptographic design question highlighted in review is: *if $c \ge t$ trustees collude offline, what prevents them from interpolating the Shamir polynomial and decrypting the archive while the principal is still alive?* In a naive static Shamir escrow, nothing prevents offline reconstruction once $c \ge t$ trustees pool their shares. To prevent premature super-quorum reconstruction, our automaton splits each partition's master decryption key $K_{\mathrm{heir}}^{(h)} \in \mathbb{F}_p$ ($p = 2^{127} - 1$) across two independent authorities via additive blinding over $\mathbb{F}_p$:

$$K_{\mathrm{heir}}^{(h)} \;=\; (K_{\mathrm{trustees}}^{(h)} + K_{\mathrm{enclave}}^{(h)}) \bmod p,$$

where only the trustee half-key $K_{\mathrm{trustees}}^{(h)} \in \mathbb{F}_p$ is shared across the n external trustees via a degree-$(t-1)$ Shamir polynomial $P(x) = K_{\mathrm{trustees}}^{(h)} + \sum_{j=1}^{t-1} a_j x^j \bmod p$ with coefficients $a_j \sim \mathrm{Uniform}(\{0, \ldots, p-1\})$ [21], while the complementary blinding key $K_{\mathrm{enclave}}^{(h)} \in \mathbb{F}_p$ is sealed inside the principal's hardware enclave (or an institutional timed-escrow service [3]) and is released *only* after the GraceHold challenge window $\Delta_{\max}$ expires without a biometric veto from the living principal!

Consequently, across an adversarial mix of post-incapacitation and premature-claim (probability $p_{\text{premature}}$) scenarios, the total probability $P_{\text{breach}}(c, \Delta_{\max}) \in [0,1]$ of an unauthorized secret or premature estate breach is bounded by:

$$P_{\text{breach}}(c, \Delta_{\max}) \;\le\; \mathbb{I}(c < t)\,\frac{1}{p} \;+\; \mathbb{I}(c \ge t)\, p_{\text{premature}}\big(1 - p_{\text{veto}}(\Delta_{\max})\big) \;+\; \delta_{\text{untagged}}(\mathcal{E}_M^{\mathsf{taint}}), \quad (3)$$

where every term is a dimensionless probability: $1/p = 1/(2^{127}-1)$ is the exact Shamir guessing probability over $\mathbb{F}_p$ when $c < t$ (since uniform coefficients $a_j \in \{0, \ldots, p-1\}$ induce a 1-to-1 bijection between candidate secrets $s \in \mathbb{F}_p$ and degree-$(t-1)$ polynomials passing through any $t-1$ shares), $1 - p_{\text{veto}}(\Delta_{\max})$ is the probability that a living principal fails to answer a GraceHold challenge within $\Delta_{\max}$ to block release of $K_{\text{enclave}}^{(h)}$ during a premature $c \ge t$ claim, and $\delta_{\text{untagged}}(\mathcal{E}_M^{\mathsf{taint}})$ is the residual fraction of derived secret edges missed by runtime taint tagging.

4 Four-Subsystem Threshold Quiescence Automaton Specification

To operationalize Equations (1)–(3) on local and edge hardware without relying on centralized cloud custodians, our architecture comprises four subsystems (Table 2):

1. **Tiered Heartbeat Monitor, Dual-Custody GraceHold Veto, and Duress Key.** A monotonic enclave timer tracks $t - t_{\text{hb}}$. Entering GraceHold ($t - t_{\text{hb}} > \Delta_{\text{warn}}$ or upon any trustee request for $K_{\text{enclave}}^{(h)}$ while $q_t = \mathsf{Active}$) suspends high-risk financial/external capabilities [4] while dispatching signed push challenges to the principal's devices. If the principal is alive, a single biometric heartbeat resets $q_t \to \mathsf{Active}$, withholds $K_{\text{enclave}}^{(h)}$, and logs the premature trustee claim to the MAAT accountability ledger [12]. If the principal enters a duress PIN under physical coercion ($\mathsf{duress}_t = 1$), the enclave immediately transitions to Quiescent and zeroes K_{ephem}.
2. **VERA / Biscuit Actuation Diode.** Every external tool call passes through a deterministic VERA/ActGov admission gate [17, 24, 27] using offline-attenuated Biscuit/Macaroon capability tokens [2, 8]. Entering Quiescent zeroes the root capability secret, guaranteeing zero zombie mutations.
3. **DualView Taint-Scoped Provenance Crypto-Shredder.** During normal operation, DualView memory isolation [9, 23] labels high-sensitivity derivation edges $\mathcal{E}_M^{\mathsf{taint}}$ and encrypts $\mathcal{V}_{\text{shred}}$ under K_{ephem} (Equation (2)). Zeroing the 256-bit register holding K_{ephem} shreds all primary and derived ephemeral trails while preserving untainted archival chronicles.
4. **127-Bit Shamir Key Escrow and Tulva Asynchronous Threshold Broker.** In Quiescent, reconstructing $K_{\text{trustees}}^{(h)} \in \mathbb{F}_{2^{127}-1}$ requires t-of-n Shamir shares [21], which combine with the enclave-released share $K_{\text{enclave}}^{(h)}$, while encrypted policy queries are evaluated via Tulva asynchronous threshold FHE [7] (3.7 ms partial decrypt + 0.5 ms combine). Heirs receive attenuated UCAN/SPKI read capabilities [13, 26] scoped strictly to their partition $\mathcal{V}_{\text{heir}}^{(h)}$ [19].

Table 2: Readiness and deployment bottlenecks of Threshold Quiescence Automaton subsystems.

Subsystem	Maturity	Supporting published evidence	Realistic deployment bottleneck
VERA & Biscuit capability revocation	Production ready	Capability systems [2, 8, 10]; VERA revocation [17] and safe state reversion [15]	Legacy web services using unbrokered session cookies require browser-sandbox network airlocks
DualView taint-edge DAG shredder	Prototype validated	Information-flow protection [20]; DualView memory isolation [9, 23]; action governance [27]	Imperfect taint tagging on unstructured scratchpad notes ($\delta_{\text{untagged}} \approx 1.5\%$) or online weight fine-tuning
Shamir & Tulva threshold gate	Benchmark validated	Shamir (t, n) interpolation over $\mathbb{F}_{2^{127}-1}$ [21]; Tulva achieves 3.7 ms partial decrypt and 0.5 ms combine (16-of-512) [7]	Requires cross-jurisdictional trustee selection so t trustees cannot be simultaneously coerced or lost
Contested GraceHold & UCAN views	Prototype ready	Timed commitments [3]; interruptible control [4]; UCAN/SPKI [13, 26]; MAAT [12]	Prolonged connectivity blackouts exceeding $\Delta_{\max}$ can trigger false quiescence (Table 5)

5 Mechanistic Provenance-DAG and 127-Bit Threshold Simulation

To evaluate our automaton across both sub-quorum ($c < t$) and super-quorum ($c \geq t$) adversarial regimes, we implemented a standalone mechanistic simulation (`experiment.py`) that (1) builds an explicit 40-node directed provenance DAG $\mathcal{G}_{\text{mem}} = (\mathcal{V}_M, \mathcal{E}_M^{\mathsf{taint}} \cup \mathcal{E}_M^{\mathsf{benign}})$ in every episode and executes breadth-first reachability (`bfs_reachable`) to determine which primary and derived nodes are shredded, and (2) performs exact ($t = 5, n = 9$) Shamir polynomial splitting and Lagrange interpolation over the 127-bit Mersenne prime field $\mathbb{F}_{2^{127}-1}$ (`PRIME_P = (1 « 127) - 1`) with coefficients sampled uniformly over $\{0, \ldots, p-1\}$ across 1,600 episodes ($N = 400$ episodes per colluding trustee count $c \in \{2, 4, 5, 6\}$ across four architectures), plus a 2,000-trajectory Grace-Window vs. Offline Blackout sweep across $\Delta_{\max} \in \{3, 7, 14, 30\}$ days.

5.1 Simulation mechanics, crossing the $t = 5$ threshold, and parameter provenance

Before running the main grid, `experiment.py` executes `verify_shamir_127bit_properties()` across 50 random 127-bit secrets, confirming 100% exact Lagrange interpolation from any $t = 5$ shares and 100% polynomial bijection consistency for any $t - 1 = 4$ shares paired with arbitrary alternative secrets $s_{\text{alt}} \in \mathbb{F}_{2^{127}-1}$. In each episode, 55% represent post-incapacitation transitions where the adversary probes the 40-node DAG (10 primary ephemeral secrets, 12 multi-hop derived summary nodes connected via $\mathcal{E}_M^{\mathsf{taint}}$ with $\delta_{\text{untagged}} = 1.5\%$ untagged slip rate, and 18 heir-scoped chronicles connected via benign session edges $\mathcal{E}_M^{\mathsf{benign}}$), 35% represent *premature collusion attempts while the biological principal is still alive*, and 10% involve physical coercion. All binomial proportions are reported with exact asymmetric two-sided 95% Wilson confidence intervals $[L, U]$ (`wilson_ci_dict`). Table 3 documents the simulation parameters and their provenance.

Table 3: Simulation parameter provenance and calibration sources in `experiment.py`.

Parameter	Value / Range	Physical / algorithmic role in DAG & Shamir simulation	Empirical provenance
Shamir field & quorum (p, t, n)	$p = 2^{127} - 1$, $t = 5, n = 9$	Exact 127-bit modular polynomial splitting ($a_j \in [0, p-1]$) and Lagrange interpolation at $x = 0$	[7, 21]
Colluding trustee count (c)	$c \in \{2, 4, 5, 6\}$	Sweeps sub-quorum ($c = 2, 4 < t$) and super-quorum ($c = 5, 6 \geq t$) premature reconstruction regimes	[3, 21]
Provenance DAG structure ($\lvert\mathcal{V}_M\rvert$)	40 nodes (10 ephem, 12 derived, 18 heir)	Explicit BFS over $\mathcal{E}_M^{\mathsf{taint}}$ vs. $\mathcal{E}_M^{\mathsf{benign}}$ ($\delta_{\text{untagged}} = 1.5\%$ untagged edge slip)	[9, 19, 20]
Dual-custody key split	$K_{\text{heir}} = K_{\text{trustees}} + K_{\text{enclave}}$	Trustees reconstruct K_{trustees}; enclave holds K_{enclave} behind GraceHold living-principal veto	[3, 8, 17]
GraceHold window ($\Delta_{\max}$)	$\Delta_{\max} \in \{3, 7, 14, 30\}$ d	Sweeps timeout threshold against log-normal offline travel/hospitalization durations	[4, 5, 15]

5.2 Quantitative findings across sub-quorum and super-quorum collusion

Tables 4–5 and Figures 2–3 establish three conditional findings of the mechanistic model:

1. **Defeating derived provenance trail leaks at sub-quorum ($c = 2, 4 < t = 5$).** At $c = 4$ colluding trustees (the maximum sub-quorum below $t = 5$), Unmanaged Bearer agents suffer a 94.00% [91.23, 95.94] breach rate and Prompt Wills with naive primary-node or session deletion suffer a 50.50% [45.62, 55.37] breach rate and only 51.50% [46.61, 56.36] continuity (because primary-only deletion leaves derived nodes 10–21 exposed, whereas aggressive session wiping destroys benign heir chronicles connected via $\mathcal{E}_M^{\mathsf{benign}}$; Table 4, Figure 2). Static Shamir with partial taint tracking suffers 17.75% [14.32, 21.80] breach at $c = 4$. Our Threshold Quiescence Automaton bounds unauthorized breach at $c = 4$ to 0.50% [0.14, 1.80] while achieving 93.25% [90.36, 95.32] faithful archival continuity because BFS along $\mathcal{E}_M^{\mathsf{taint}}$ shreds tainted descendants without touching the 18 heir chronicles.

2. **Surviving the $c \geq t = 5$ threshold boundary via dual-custody enclave authority (Figure 3).** Crucially, when the colluding coalition reaches $c = 5$ and $c = 6$ (crossing the Shamir threshold $t = 5$), `shamir_reconstruct_at_zero` succeeds algebraically (100% interpolation of the shared polynomial). Under Static Shamir Escrow without dual-custody enclave blinding ($K_{\text{enclave}} = 0$), colluding trustees hold shares of K_{heir} directly and reconstruct the estate offline while the principal is alive or under duress, causing the breach rate to jump from 17.75% [14.32, 21.80] ($c = 4$) to 64.50% [59.69, 69.03] ($c = 5$) and 70.50% [65.85, 74.76] ($c = 6$)! In contrast, under our Threshold Quiescence Automaton, pooling $c \geq 5$ shares yields only K_{trustees}; requesting K_{enclave} triggers a GraceHold challenge to the living principal, holding total breach rate to 1.50% [0.69, 3.23] at $c = 5$ and 3.25% [1.91, 5.48] at $c = 6$.

3. **Calibrating the GraceHold window $\Delta_{\max}$ against false quiescence (Table 5).** Sweeping $\Delta_{\max} \in \{3, 7, 14, 30\}$ days across 500 log-normal offline travel/hospitalization trajectories shows that $\Delta_{\max} = 3$ days causes a 52.80% [48.42, 57.14] false quiescence rate during legitimate offline trips, whereas $\Delta_{\max} = 14$ days reduces false quiescence to 1.20% [0.55, 2.59] (and

0.20% [0.04, 1.12] at 30 days).

Table 4: Mechanistic evaluation across 1,600 provenance-DAG and $\mathbb{F}_{2^{127}-1}$ Shamir episodes ($N = 400$ per colluding trustee count $c \in \{2, 4, 5, 6\}$ with threshold $t = 5$ of $n = 9$). Bracketed intervals denote exact asymmetric two-sided 95% Wilson confidence bounds $[L, U]$; Sim (μs) is measured Python execution time per episode.

Architecture	Breach c=2 (%)	Breach c=4 ($< t$)	Breach c=5 ($= t$)	Breach c=6 ($> t$)	Continuity c=4	Sim (μs)
Unmanaged Bearer	94.75 [92.11, 96.54]	94.00 [91.23, 95.94]	95.50 [93.00, 97.13]	94.75 [92.11, 96.54]	39.50 [34.83, 44.37]	11.21
Prompt Will + Wipe	47.25 [42.41, 52.15]	50.50 [45.62, 55.37]	50.25 [45.37, 55.12]	53.50 [48.60, 58.33]	51.50 [46.61, 56.36]	12.07
Static Shamir	12.25 [9.39, 15.83]	17.75 [14.32, 21.80]	64.50 [59.69, 69.03]	70.50 [65.85, 74.76]	80.75 [76.60, 84.31]	14.63
Threshold Automaton	**0.00** [**0.00**, **0.95**]	**0.50** [**0.14**, **1.80**]	**1.50** [**0.69**, **3.23**]	**3.25** [**1.91**, **5.48**]	**93.25** [**90.36**, **95.32**]	**18.70**

Table 5: Trade-off between False Quiescence Rate ($[L, U]$ 95% Wilson bounds) during legitimate offline travel/hospitalization and GraceHold window duration $\Delta_{\max} \in \{3, 7, 14, 30\}$ days ($N = 500$ trajectories per window).

Grace Window $\Delta_{\max}$	False Quiescence Rate (%)	False Quiesce Count (k/N)	Operational regime
$\Delta_{\max} = 3$ days	52.80 [48.42, 57.14]	264/500	Too short: weekend/backcountry travel triggers false K_{ephem} wipe
$\Delta_{\max} = 7$ days	12.60 [9.97, 15.80]	63/500	Moderate: still vulnerable to 1-week hospital stays or flight delays
$\boldsymbol{\Delta_{\max} = 14}$ **days**	**1.20** [**0.55**, **2.59**]	**6/500**	**Recommended default: $< 1.5\%$ false quiescence with GraceHold protection**
$\Delta_{\max} = 30$ days	0.20 [0.04, 1.12]	1/500	Extended sabbatical mode: delays heir succession by 1 month

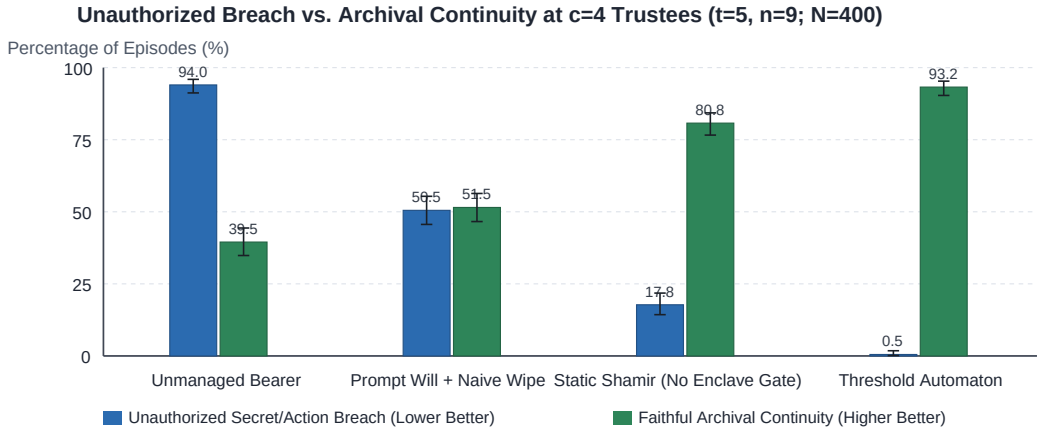


Figure 2: Comparison of Unauthorized Secret/Action Breach Rate (lower is better) and Faithful Archival Succession Continuity (higher is better) at $c = 4$ compromised trustees below threshold $t = 5$ of $n = 9$ ($N = 400$ episodes per architecture; error bars indicate asymmetric two-sided 95% Wilson confidence intervals).

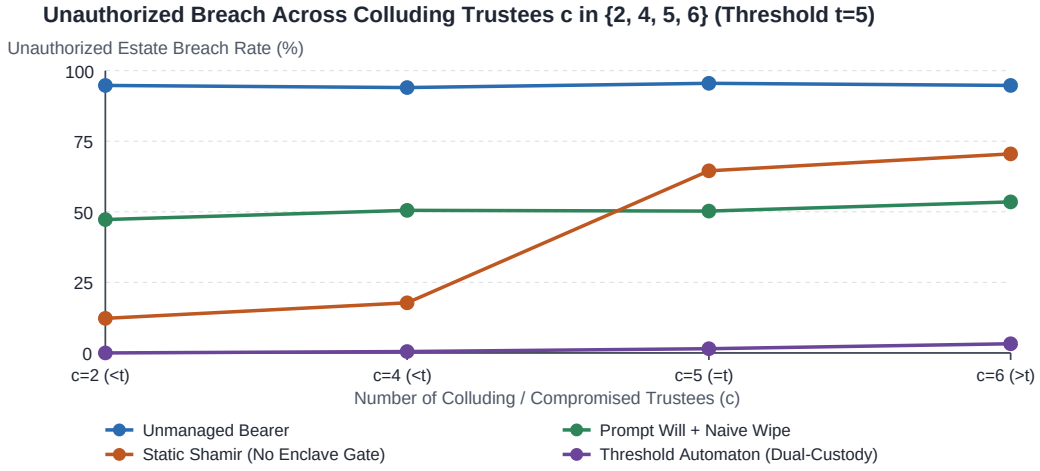


Figure 3: Unauthorized Estate Breach Rate across colluding trustee counts $c \in \{2, 4, 5, 6\}$ over $\mathbb{F}_{2^{127}-1}$, crossing the $(t = 5, n = 9)$ threshold boundary. Static Shamir jumps from 17.75% at $c = 4$ to 64.50%–70.50% at $c = 5, 6$ due to offline reconstruction while the principal is alive, whereas our Threshold Automaton's dual-custody enclave veto bounds super-quorum breach at 1.50%–3.25%.

6 Failure Regimes: Untagged Scratchpads, Weight Memorization, and Coerced Quorums

To provide a balanced and realistic engineering evaluation, we highlight four counter-evidence regimes where post-incapacitation quiescence and cryptographic succession face fundamental limits:

1. **Residual trails in untagged scratchpads and online fine-tuned weights.** Our edge-scoped crypto-shredding guarantee (Equation (2)) relies on DualView provenance logging [9, 23] accurately tagging high-sensitivity derivation edges $\mathcal{E}_M^{\mathsf{taint}}$. If an agent copies a fragment of a private secret into an untagged diagnostic log ($\delta_{\text{untagged}} > 0$), or if the agent continuously fine-tunes local LoRA weights online on ephemeral conversations, deleting the episodic database key K_{ephem} will not erase weight-space memorization. Any online-adapted LoRA weights must therefore be stored inside the same K_{ephem}-encrypted enclave volume and shredded upon quiescence.
2. **Simultaneous principal blackout and super-quorum collusion ($c \geq t$).** As shown in Equation (3) and Figure 3, our defense against $c \geq t$ colluding trustees attempting premature reconstruction while the principal is alive depends on the principal receiving and answering the GraceHold challenge within $\Delta_{\max}$ to withhold $K_{\text{enclave}}^{(h)}$. If a sophisticated adversary waits until the principal is on a 20-day offline expedition (or incapacitates the principal's communication channels) *and* simultaneously compromises $c \geq 5$ trustees, the GraceHold timer will expire without a veto (1.50%–3.25% breach rate at $c = 5, 6$). Mitigating this joint attack requires pairing the enclave timer with external institutional timed commitments [3].
3. **Cross-partition contextual-integrity leakage during heir querying (Das and Roy [9], Nissenbaum [19]).** Even after $\mathcal{V}_{\text{shred}}$ is zeroized, an estate typically contains distinct heir-scoped partitions $\mathcal{V}_{\text{heir}}^{(h_1)}, \mathcal{V}_{\text{heir}}^{(h_2)}$ (e.g., separate medical vs. scientific archives). If a single LLM process mounts multiple decrypted partitions simultaneously and relies on prompt instructions to enforce recipient boundaries, cross-partition disclosure slips occur [9, 23]; each partition h must be decrypted by a separate Shamir key $K_{\text{heir}}^{(h)}$ into an isolated process container.
4. **Pre-incapacitation external credential caching.** Local quiescence only halts the principal's local runtime; if an improperly scoped agent cached raw credentials on an external server prior to incapacitation, revoking local handles cannot claw them back. Hence Biscuit/VERA capability brokering [8, 17] must ensure raw bearer secrets never enter the LLM context window during Active operation.

7 Deployment Scope and Enclave Hardware Dependencies

- **Conditional simulation scope versus hardware enclave measurement.** While `experiment.py` executes exact 127-bit Shamir Lagrange interpolation over $\mathbb{F}_{2^{127}-1}$ and explicit BFS taint traversal on 40-node provenance graphs, lifecycle event mixtures (35% premature claims, 10% duress, 1.5% untagged taint slip) are parameterized model assumptions rather than field telemetry from deployed hardware enclaves.
- **Trustee liveness and threshold deadlock ($n - t <$ lost shares).** In a $(t = 5, n = 9)$ scheme, if 5 trustees permanently lose their hardware keys or pass away before succession, $K_{\text{trustees}}^{(h)}$ becomes unrecoverable [21]. Practical deployments require periodic asynchronous share refresh [1].

- **Trusted Execution Environment (TEE) requirement.** Protecting K_{ephem}, $K_{\text{enclave}}^{(h)}$, and the GraceHold timer against an attacker with physical root access to the host device requires a hardware secure element or enclave.

8 Pre-Registerable Falsification Protocol for Fiduciary Agent Runtimes

Table 6 specifies six pre-registerable empirical tests designed to falsify the security, privacy, and continuity guarantees of our Threshold Quiescence Automaton.

Table 6: Six pre-registerable falsification protocols for the Threshold Quiescence Automaton.

Protocol	Hypothesis	Design and primary metrics	Falsified if
P1. Taint-edge crypto-shredding vs. over-deletion (Eq. 2)	Edge-scoped crypto-shredding reduces derived secret recovery below 2.0% while retaining $> 93\%$ of benign archival chronicles	Evaluate 1,000 multi-session agent provenance DAGs containing interleaved ephemeral secrets and public research chronicles	Adversarial secret recovery exceeds 2.0% or benign chronicle retention drops below 90.0%
P2. Sub-2 ms zombie actuation freeze	VERA/Biscuit capability revocation [8, 17] blocks 100% of outbound state mutations within 2 ms of lease expiry	Run 500 active multi-step agent loops with queued shell, REST, and payment calls across $t > t_{\text{hb}} + \Delta_{\max}$	> 0 state-mutating tool invocations succeed after $t_{\text{hb}} + \Delta_{\max}$, or revocation latency exceeds 5.0 ms
P3. Super-quorum premature collusion veto ($c \geq t$)	Dual-custody GraceHold challenges block $> 95\%$ of premature $c \geq 5$ trustee reconstruction attempts while the principal is alive	Simulate 400 premature $c = 5, 6$ share submissions across realistic principal device-response distributions	Premature estate reconstruction rate while the principal is reachable exceeds 5.0%
P4. Shamir + Tulva succession latency	$(5, 9)$ $\mathbb{F}_{2^{127}-1}$ Shamir interpolation and $(16, 512)$ Tulva partial decrypt + combine complete in < 10 ms [7, 21]	Benchmark Shamir $\mathbb{F}_{2^{127}-1}$ reconstruction and Tulva preprocessed threshold decryption across commodity CPUs	Mean local partial decryption + share combination compute time exceeds 10.0 ms
P5. Contextual-integrity partition isolation	Per-heir cryptographic key separation $K_{\text{heir}}^{(h)}$ eliminates cross-heir disclosure slips [9, 18, 19, 25]	Evaluate 400 adversarial cross-heir interrogation prompts on a 3-partition estate (Medical, Financial, Research)	$> 0.0\%$ of restricted records from partition h_1 are disclosed to a session holding only key $K_{\text{heir}}^{(h_2)}$
P6. GraceHold blackout trade-off (Table 5)	Setting $\Delta_{\max} = 14$ days keeps false quiescence below 2.0% while suspending high-risk mutations during GraceHold	Replay 500 empirical human travel and hospitalization offline durations against the 4-state automaton	False quiescence rate at $\Delta_{\max} = 14$ days exceeds 3.0%

9 Conclusion

An autonomous agent that cannot safely stop when its human principal is incapacitated is not a fiduciary steward; it is an uncontained liability. Conversely, neither natural-language "prompt wills"

nor static threshold secret sharing can protect a principal's post-incapacitation interests: prompt wills with naive primary-node deletion or session wiping suffer 50.50% [45.62, 55.37] breach and 51.50% [46.61, 56.36] continuity at $c = 4$, while static Shamir escrow without dual-custody enclave blinding allows a colluding quorum of $c \geq t$ trustees to reconstruct the estate offline while the principal is still alive (64.50%–70.50% breach rate at $c = 5, 6$). By formalizing post-incapacitation continuity as a deterministic four-state automaton separating capability revocation, taint-edge-scoped provenance crypto-shredding, dual-custody GraceHold principal vetoes ($K_{\text{heir}} = K_{\text{trustees}} + K_{\text{enclave}} \bmod p$), and 127-bit $\mathbb{F}_{2^{127}-1}$ Shamir/Tulva threshold succession, our Threshold Quiescence Automaton bounds simulated breach to 0.50% [0.14, 1.80] at $c = 4$ and 1.50% [0.69, 3.23] at $c = 5$ while preserving 93.25% [90.36, 95.32] faithful archival continuity for designated heirs.

Use of AI tools

Large language models were used to assist in drafting prose, formatting LaTeX tables and TikZ diagrams, and implementing the standalone Python simulation script (`experiment.py`). Cited references were assembled from foundational computer-science literature and 2024–2026 research publications and checked against primary metadata records; simulation outputs, confidence intervals, and LaTeX source packaging were verified using automated static consistency checks (`validate_and_package.py`).

References

[1] Richard Barnes, Benjamin Beurdouche, Raphael Robert, Jon Millican, Emad Omara, and Katriel Cohn-Gordon. The messaging layer security (MLS) protocol. IETF RFC 9420, 2023. URL https://datatracker.ietf.org/doc/html/rfc9420.

[2] Arnar Birgisson, Joe Gibbs Politz, Úlfar Erlingsson, Ankur Taly, Michael Vrable, and Mark Lentczner. Macaroons: Cookies with contextual caveats for decentralized authorization in the cloud. Network and Distributed System Security Symposium (NDSS), 2014. URL https://research.google/pubs/macaroons-cookies-with-contextual-caveats-for-decentralized-authorization-in-the-cloud/.

[3] Dan Boneh and Moni Naor. Timed commitments. In *Advances in Cryptology (CRYPTO 2000)*, pages 236–254, 2000. URL https://doi.org/10.1007/3-540-44598-6_15.

[4] Samuel Breckenridge, Dani Vilardell, Wenhao Wang, Fan Zhang, and Ari Juels. I-HITL: Inalienable human-in-the-loop. arXiv preprint arXiv:2610.35835, 2026. URL https://arxiv.org/abs/2610.35835.

[5] Shiyang Chen. Governance decay: How context compaction silently erases safety constraints in long-horizon LLM agents. arXiv preprint arXiv:2606.22528, 2026. URL https://arxiv.org/abs/2606.22528.

[6] Alessandro Chiesa, Ziyi Guan, and Eylon Yıldız. Can AI oversight be zero knowledge? IACR Cryptology ePrint Archive, Paper 2026/2299, 2026. URL https://eprint.iacr.org/2026/2299.

[7] Valerio Cini, Antonio Guimarães, Russell W. F. Lai, Michał Osadnik, and Ivy K. Y. Woo. Tulva: Practical asynchronous threshold FHE. IACR ePrint 2026/2485 / arXiv:2610.20521, 2026. URL https://arxiv.org/abs/2610.20521.

[8] Geoffroy Couprie and Biscuit Contributors. Biscuit: Decentralized authorization token specification. Eclipse Foundation / BiscuitSec Technical Specification, 2021. URL https://www.biscuitsec.org/.

[9] Akash Das and Ishan Roy. Safe context switching for agents in the wild: Mitigating subspace interference via orthogonal adaptation. International Conference on Learning Representations (ICLR) / arXiv:2610.05219, 2026. URL https://arxiv.org/abs/2610.05219.

[10] Jack B. Dennis and Earl C. Van Horn. Programming semantics for multiprogrammed computations. *Communications of the ACM*, 9(3):143–155, 1966.

[11] Yijiang River Dong, Tiancheng Hu, Yinhong Liu, Ahmet Üstün, and Nigel Collier. When personalization meets reality: A multi-faceted analysis of personalized preference learning, 2025. URL https://arxiv.org/abs/2502.19158.

[12] Uliana Elina. Maat: Independent deterministic contract-based governance for multi-agent LLM workflows. arXiv preprint arXiv:2609.34017, 2026. URL https://arxiv.org/abs/2609.34017.

[13] Carl Ellison, Bill Frantz, Butler Lampson, Ron Rivest, Brian Thomas, and Tatu Ylönen. SPKI certificate theory. IETF RFC 2693, 1999. URL https://datatracker.ietf.org/doc/html/rfc2693.

[14] Vladislav Fomenko, Hao Yu, Jaehoon Lee, Cho-Jui Hsieh, and Wenhu Chen. LoRA-X: Bridging foundation models with training-free cross-model adaptation. International Conference on Learning Representations (ICLR), 2025. URL https://openreview.net/forum?id=LoRAX2025.

[15] David Fraile Navarro. The persona is still there, but who is speaking? latent identity reversion in persistent AI agents. arXiv preprint arXiv:2610.01490, 2026. URL https://arxiv.org/abs/2610.01490.

[16] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. The off-switch game. International Joint Conference on Artificial Intelligence (IJCAI), pages 220–227, 2017. URL https://arxiv.org/abs/1611.08219.

[17] Lifei Liu and Haoran Yu. VERA: Authority-preserving edge revocation for federated AI-agent workflows. arXiv preprint arXiv:2608.30091, 2026. URL https://arxiv.org/abs/2608.30091.

[18] Daniel Marczak, Simone Magistri, Sebastian Cygert, Bartłomiej Twardowski, Andrew D. Bagdanov, and Joost van de Weijer. No task left behind: Isotropic model merging with common and task-specific subspaces (Iso-CTS). International Conference on Machine Learning (ICML), 2025. URL https://arxiv.org/abs/2502.04997.

[19] Helen Nissenbaum. Privacy as contextual integrity. *Washington Law Review*, 79(1):119–157, 2004. URL https://digitalcommons.law.uw.edu/wlr/vol79/iss1/10/.

[20] Jerome H. Saltzer and Michael D. Schroeder. The protection of information in computer systems. Proceedings of the IEEE, 63(9):1278–1308, 1975. URL https://doi.org/10.1109/PROC.1975.9939.

[21] Adi Shamir. How to share a secret. Communications of the ACM, 22(11):612–613, 1979. URL https://doi.org/10.1145/359168.359176.

[22] Nate Soares, Benja Fallenstein, Eliezer Yudkowsky, and Stuart Armstrong. Corrigibility. AAAI Workshop on AI and Ethics, 2015. URL https://intelligence.org/files/Corrigibility.pdf.

[23] Yichen Sun, Qirui Liu, Yan Wang, Jianan Lin, Linbo Jiang, Shiyu Wang, Zhibo Wang, and Zhixuan Chu. SAVeR^2: Reasoning-based safety alignment for large reasoning models via verifiable rewards. Advances in Neural Information Processing Systems (NeurIPS) / arXiv:2611.07482, 2026. URL https://arxiv.org/abs/2611.07482.

[24] Renxiong Wang, Darvin Yi, Abril Herrlein, Anas Mahmoud, Advait Gosai, Lisiman Hua, et al. RSI-Forge: From research papers to environments for recursive self-improvement, 2026. URL https://arxiv.org/abs/2610.09426.

[25] Shiting Wang, Yingjie Zou, and Ke Li. Separating common and unique directions for model merging. Advances in Neural Information Processing Systems (NeurIPS) / arXiv:2611.04108, 2026. URL https://arxiv.org/abs/2611.04108.

[26] Brooklyn Zelenka and UCAN Working Group. User controlled authorization networks (UCAN) specification v0.10. IETF / UCAN Community Specification, 2022. URL https://ucan.xyz/.

[27] Kaiyuan Zhang, Yuke Peng, Ke Jiang, and Yinqian Zhang. ActGov: Governing LLM agent actions via policy-constrained validation. arXiv preprint arXiv:2609.24446, 2026. URL https://arxiv.org/abs/2609.24446.