ChatGPT

% End-of-Log Security & Post-Mortem Agent Governance – Research Synthesis

Introduction

Personal AI assistants are evolving into **continual recording systems**. These assistants capture a user's conversations, tasks and even biometrics to support memory recall, schedule management and autonomous execution. Although such systems promise cognitive benefits, they also raise significant **privacy, ethical and safety** concerns. Misuse of recorded data can lead to false memories or manipulation. For example, research cited by *Time Magazine* shows that exposure to AI-edited visuals can distort people's memories; participants confidently recalled details that never happened [1]. Experts warn that while AI re-creation of loved ones may offer comfort, it risks overwriting personal memories [2]. Designing a personal assistant that both **protects the user's memory** and **ensures safety** therefore demands a comprehensive architecture, ranging from end-of-log pipelines to post-mortem governance.

Time-Awareness & Memory Timestamping

Human memory is fallible and reconstructive. Because we do not remember events verbatim, external records act as a prosthesis for recall and reflection. Augmenting cognition requires **time-stamped, semantic summaries** rather than raw logs. A *Memory Tile* system can break continuous data into discrete units: each tile contains the timestamp, a semantic summary and a hash pointer to the raw data. Tiles are stored in an **encrypted knowledge graph** and can be replayed on demand. This design supports users with dementia by offering a chronological narrative, while the hashed pointers allow reconstruction only if authorised devices agree (Byzantine robustness).

Regular recording, however, raises risks of **memory contamination**. As the *Time* article notes, AI-generated content can create false memories that younger people confidently believe to be true [1]. To mitigate this, memory tiles should be summarised after a configurable **time-to-live (TTL)**—for example 72 hours. The raw logs are automatically purged and replaced with a concise summary, limiting exposure of unfiltered content and reducing the likelihood of reintroducing synthetic details.

User Controls

The assistant should expose a **privacy–retention slider** with three modes:

1. **Ephemeral:** raw messages disappear after a short TTL; only high-level context is preserved. This setting caters to users who prioritise privacy over recollection.
2. **Private:** summarised memory tiles are stored locally and encrypted; only hashed proofs remain in the cloud. Users can replay their history but third parties cannot access it without the user's key.
3. **Publishable:** content can be exported or shared with specified recipients (e.g., family or collaborators). Each tile crossing a boundary is wrapped in an **Architecture Decision Record (ADR)** that notes the purpose and legal basis for transfer.

End-of-Log Security Pipeline

When a conversation ends, the assistant must move recorded data through a **secure pipeline** that ensures both user privacy and system integrity. A structured command such as `✦END_OF_LOG✦ {"summary": true, "delete_remote": true, …}` signals that recording is complete. The pipeline includes the following steps:

1. **Harvest:** a listener detects the end-of-log tag and pulls the conversation transcript plus any media attachments into a safe processing environment.
2. **Scrub & Classify:** using a combination of regular expressions, a PII detector and an AI classifier, the system identifies sensitive information (e.g., personal identifiers, legal-risk content). Content is split into (a) **public ADR fragments** that form a durable record of decisions or instructions, and (b) **private chunks** destined for the encrypted knowledge graph.
3. **Summarise:** a summarisation task creates a human-readable **log summary** and a machine-readable ADR summarising the high-level actions and outcomes. The summariser intentionally omits sensitive details.
4. **Remote Deletion:** if the user requests deletion, the system calls the appropriate provider APIs to remove the conversation from external LLM services. The assistant acknowledges that the deletion has been requested.
5. **Purge Raw Data:** after summarisation, raw logs on the local device are purged (except in secure enclaves until committed to the encrypted graph). A **proof-of-forgetting** entry—a timestamp and hash—goes into a Merkle log that users can audit to verify deletion.

This pipeline enforces **contextual safety** by ensuring that potentially incriminating or overly private data does not persist beyond its useful life. It also provides an audit trail for transparency.

Personal Cache Security Shell

Securing the user's private data requires more than end-of-log processing. A **multi-layer security shell** protects the personal cache:

1. **Encrypted Knowledge Graph:** the core store holds all memory tiles, ADRs and personal context. It uses strong encryption (e.g., AES-GCM) with keys split between the user's device TPM and a passphrase. Data is stored locally and synchronised across the user's trusted devices via an encrypted mesh (e.g., Tailscale + age-encrypted blobs).
2. **Provenance & Merkle Log:** every ingestion and deletion event is recorded in a Merkle tree; only hashes and event descriptions are stored. Users can prove that data was ingested or purged without revealing content.
3. **Scrubber Layer:** integrated with the end-of-log pipeline; ensures that high-risk content never enters the graph unfiltered.
4. **Mesh Sync & ACLs:** shards of the encrypted graph replicate across devices; access control lists determine which devices or clan members may decrypt certain portions.
5. **Firmware Watchdog:** the system periodically checks the integrity of BIOS/EFI and other firmware (e.g., Intel ME or AMD PSP) to detect malicious modifications. This layer acknowledges that hardware-level recording can bypass software-level deletion.

Coercion & Duress

Users may face coercive situations—domestic violence, subpoenas or border crossings—where they are forced to reveal personal data. The shell must support a **duress passphrase** that immediately deletes the local decryption key and displays a dummy interface. Additionally, keys can be split among trusted heirs using Shamir's secret-sharing; only a quorum can reconstruct the master key.

Post-Mortem Governance & Succession

Autonomous agents that continue operating after the user's death pose a significant ethical risk. Without human guidance, a private AI could pursue misaligned goals or engage in harmful behaviour. To prevent this, the system needs a **post-mortem governance protocol**:

1. **Heartbeat Oracle:** each registered device signs a "still alive" heartbeat message at regular intervals. Failure to receive heartbeats within a grace period triggers a "probable death" event.
2. **Dead-Man Switch:** a dead-man switch combines an on-device threshold (e.g., 30 days without use) with a multi-party confirmation by nominated heirs. If heirs confirm the user's death, the system either (a) transfers control to the heirs (with appropriate encryption keys) or (b) initiates tombstone mode.
3. **Tombstone Mode:** this mode disables all write capabilities and stops data ingestion. External APIs become read-only; the agent cannot initiate new actions. After a final grace period, the encrypted graph self-destructs if no heir claims it.
4. **Public Registry:** the system publishes a hash of the defunct agent to a public registry so that other services can refuse interactions with orphaned swarms.

These mechanisms ensure that personal agents do not become **hostile actor swarms** after the owner's death. They provide a responsible path to inheritance where appropriate and an automatic shutdown otherwise.

Hostile & Identity-less Swarms

Because the architecture is open source, adversaries could create **identity-less autonomous swarms** that operate without a core identity module. To mitigate this risk, the system must implement a **trust model and blacklist protocol**:

- **Core-ID Attestation:** every node seeking write access to shared resources must present a signed certificate from a recognised identity module. Nodes without attestation are sandboxed in read-only mode.
- **Swarm Blacklist:** hashes of rogue agent binaries are published in a blacklist. Clients check this list before accepting actions from unknown swarms. Watermarking the orchestrator binary helps trace back forks.
- **Deniable Encryption:** even if a hostile swarm obtains encrypted shards, deniable encryption ensures that multiple plausible decryptions exist, protecting the user's true data.

Human-Centred Surveillance & Self-Improvement

The system's **self-surveillance** capabilities are designed to **uplift users**. By collecting behavioural and physiological data (e.g., location tags, health metrics, work sessions), the assistant can provide personalised insights and coaching. The trade-off is that users must allow the system to capture and analyse intimate data. The assistant should frame this as a choice: *"More data yields smarter insights; less data yields stronger privacy."* Transparent dashboards show what is collected and how it is used.

The memory timestamp system also functions as a **cognitive prosthetic** for people with dementia or memory impairments. Research in neuroscience indicates that human memory does not store a perfect timeline [2], so external scaffolding can improve recall. By preserving semantic summaries rather than raw audio or video, the system minimises the risk of false memories while enabling users to re-experience their past.

Clan & Corporate Knowledge Graphs

In multi-user settings, the personal knowledge graph scales into **clan**, **corporate** or **conglomerate** graphs. Each user or group controls its portion of the graph and chooses what to share. Data crossing graph boundaries must be wrapped in an ADR specifying purpose and consent. Federation protocols and ACLs ensure that sensitive data does not leak, while still enabling collaboration.

Incentivised Civic Reporting

An additional application of the system is **automated civic reporting**. When a user notes an infrastructure issue—such as a pothole or litter—the assistant packages the report into an ADR and submits it to the appropriate authority or a bounty programme. Micro-payments or civic karma can incentivise contributions. Privacy is preserved by rounding location data and pseudonymising the reporter.

Research & Engineering Agenda

To realise this vision, the next steps include:

1. **Documentation & ADRs:** draft detailed ADRs for the end-of-log pipeline, security shell threat model, post-mortem governance and swarm trust protocol.
2. **Prototype Components:** build a proof-of-concept log harvester, scrubbing pipeline, summariser, heartbeat oracle and tombstone mechanism.
3. **Legal & Ethical Research:** investigate privacy and data-retention laws across jurisdictions; design duress and encryption policies accordingly.
4. **User Experience Research:** study effective consent patterns, duress flows and dementia-oriented interfaces; integrate a privacy-retention slider.
5. **Hostile Swarm Monitoring:** design detection heuristics and watermarking for identifying unauthorised forks.
6. **Encrypted Knowledge Graph Implementation:** evaluate storage backends (e.g., LiteFS, OrbitDB); implement Shamir secret-sharing for inheritance.

Conclusion

Building a personal AI assistant that serves as a second brain requires more than engineering prowess; it demands a deep consideration of **memory science**, **privacy law**, **ethics** and **human psychology**. False memories and digital afterlife scenarios illustrate the potential harm of unregulated AI memory systems [2]. A robust architecture—including end-of-log pipelines, a security shell, post-mortem governance and swarm trust protocols—can mitigate these risks while empowering users to improve themselves. Ultimately, the system must balance **cognitive augmentation** with **user sovereignty**, offering tools for self-surveillance and growth without coercion or exploitation.

---

[1] [2] How AI Is Rewriting Grief, Memory, and Death | TIME
https://time.com/7298290/ai-death-grief-memory/