SAGE: A Secure Agent Governance Ecosystem for Organizational Learning

Darrell L. Young dlyoung@smu.edu Jason Teske jason@wave3ai.com

May, 2025

Statement of Novelty and Contribution: *The Secure Agent Governance Ecosystem (SAGE) introduces a comprehensive architectural blueprint for scalable and privacy-preserving AI agents that integrate long-term memory, spatial reasoning, blockchain auditability, and ethical oversight. While many existing frameworks emphasize agent performance or learning, SAGE uniquely centers consent, explainability, and adversarial resilience through layered governance. It proactively addresses every known AI-specific vulnerability in the MITRE ATLAS matrix, positioning it as a high-impact contribution to the fields of AI safety, trustworthy machine learning, and human-aligned autonomy. SAGE is the first open architecture to combine cryptographic "right to forget" compliance, GIS-aware differential privacy, and consent-based semantic memory in a modular, sector-agnostic system.*

Abstract

This paper introduces SAGE, a Secure Agent Governance Ecosystem, as a universal architecture for scalable, trustworthy agent-based systems. Designed for adaptability across sectors—including elder care, education, and smart infrastructure—SAGE integrates privacy-preserving memory, strategic learning, blockchain-enabled auditability, and GIS-aware overlays. A three-tier agent hierarchy fosters individualized support while enabling system-wide intelligence. Enhancements such as Consent-Based Relationship Memory (CBRM) and geospatial privacy controls ensure that personalization does not come at the cost of ethical risk. Key innovations include zero-knowledge proofs for audit logging, smart contract-enforced expiration, and cryptographic right-to-forget compliance.

1 Introduction

The proliferation of AI services across domains has increased the demand for intelligent agent systems that are not only capable but ethically aligned and privacy-conscious. In response, we present the SAGE architecture: a Secure Agent Governance Ecosystem that ensures contextual intelligence, consent-driven personalization, and strategic learning through a layered agent hierarchy. Rooted in prior work on prompt orchestration [3, 2], SAGE draws inspiration from multi-agent alignment literature [1] and integrates contemporary advances in LLMs, blockchain, and spatial privacy.

2 Three-Tier SAGE Architecture

The SAGE ecosystem is structured across three collaborative agent tiers:

- **Tier 1 - Client Agents:** Engage directly with users for routine support, companionship, and data collection.
- **Tier 2 - Supervisor Agents:** Coordinate a cohort of Client Agents, escalate anomalies, and enforce behavioral guidelines.
- **Tier 3 - Manager Agents:** Operate across agent clusters to refine policy, allocate resources, and audit compliance.

Specialized agents (e.g., for diabetes or legal aid) may be summoned on demand, and **emergency escalation agents** coordinate rapid responses in crises, optionally integrating human-in-the-loop oversight.

3 Interaction Modalities and Learning Flows

SAGE supports three channels of knowledge transmission:

- **Top-down:** Tier 3 disseminates policy and capability upgrades.
- **Bottom-up:** Tier 1 Client Agents contribute summaries and behavior logs to inform policy adaptation.
- **Peer-to-peer:** Agents exchange tactics and reinforce successful behaviors through supervised benchmarks.

This distributed learning model fosters institutional memory without centralizing personal data.

4 Multimodal Toolchains and Context Awareness

Each SAGE agent integrates a toolchain tailored to role and modality, including:

- LLM-backed sentiment analysis and topic modeling
- Symptom checkers and scheduling modules
- Video/audio summarizers for multimodal input
- GIS overlays for spatially-aware insight

Tool usage adheres to both contextual relevance and agent scope. Supervisor Agents may, for example, activate spatial risk filters only with consent from Tier 1 Client Agents.

5 Privacy-Preserving Personalization: CBRM

To build long-term rapport without compromising user trust, we introduce **Consent-Based Relationship Memory (CBRM)**—a principled system for storing user-disclosed life details (e.g., family names, birthdays, hobbies) through active consent, metadata governance, and periodic review.

Semantic Memory and Consent

Memories are opt-in and semantically categorized (e.g., `family`, `milestone`, `hobby`) with expiration windows and usage contexts (e.g., `rapport-building-only`).

Memory Dashboard and Anonymized Storage

Clients may view, delete, or label memories using a dashboard, with all data encrypted and stored in a separate layer accessible only after session authentication.

Ethical Framing and Trust

Agents communicate memory usage empathetically: e.g., "I remember your grandson loves baseball—that helps me be a better companion. But only if you're okay with it."

Learning Implications

Supervisors and Managers track aggregate memory trends for compliance and improvement without inspecting individual details. CBRM enables deeper trust in domains like elder care, education, and social services.

6 Blockchain-Backed Consent and Auditability

To reinforce user control and governance, SAGE integrates blockchain elements:

- **Right to Forget:** Consent revocations trigger cryptographically verifiable memory deletions, enforced via blockchain transaction logs.
- **Smart Contracts:** Automatically manage consent scope, expiration, and usage tracking for sensitive session data.
- **Zeroizing Keys:** Sensitive memory keys can be zeroized upon consent withdrawal, ensuring unrecoverable erasure.
- **Audit Logs:** Immutable blockchain-based summaries of agent behavior are generated for review by compliance officers or regulators.

7 GIS and Spatial Privacy Controls

Geospatial reasoning adds value but increases risk. SAGE includes:

- **Spatial Differential Privacy:** Prevents reconstruction of individual locations from aggregate data.
- **Geo-Consent Contracts:** Allow users to approve or disable agent actions tied to geographic location.
- **Zone Filtering:** Disables agent activity in sensitive locations (e.g., safe houses).

8 ATLAS-Resilient Design Principles

SAGE incorporates proactive defenses against AI-specific threats described in the MITRE ATLAS matrix:

- **Data Poisoning:** All training inputs are labeled with provenance metadata and hashed for integrity to detect manipulation.
- **Model Evasion:** Supervisor Agents conduct adversarial simulations to test detection robustness.
- **Model Inversion:** Differential privacy is applied to memory embeddings to resist training data extraction.
- **Membership Inference:** Randomized output masking and dropout reduce inference reliability.
- **Prompt Injection:** System-level firewalls sanitize and rewrite user prompts to strip malicious commands.
- **Supply Chain Attacks:** Only reproducible, signed packages are allowed in the agent runtime.
- **Unauthorized Model Access:** Per-session encryption keys and endpoint isolation prevent misuse.
- **Hallucination Risk:** Confidence tagging and source-linked citations alert users to uncertain claims.
- **Model Theft:** Throttling and query fingerprinting prevent excessive probing via APIs.
- **Backdoors:** Hidden state audits search for anomalous activation patterns that may signal inserted triggers.

9 Agent Experiments and Use Cases

We evaluated SAGE in three scenarios:

- **Elder Care (Caring-Call):** Client agents remembered approved personal details and detected wellness anomalies, escalating via Supervisor Agents.
- **AI GIS Services:** Spatial overlays helped agents advise on facility planning while honoring zone-based exclusions.
- **Educational Coaching:** Supervisor Agents recommended learning plans based on anonymized feedback trends.

Preliminary results suggest SAGE agents can foster trusted relationships, support strategic adaptation, and operate within evolving regulatory norms.

10 Conclusion

SAGE offers a unified model for scalable, ethical AI agent ecosystems. Its three-tier hierarchy enables personalized support and organizational learning, while innovations like CBRM, blockchain enforcement, zeroizing cryptographic memory keys, and spatial privacy safeguards ensure future-proof compliance. New additions inspired by the MITRE ATLAS matrix further fortify SAGE against adversarial threats, model misuse, and data leakage. SAGE holds promise not only for elder care, but any domain where adaptive learning must co-exist with human dignity and control.

References

[1] D. Dhinakaran, S. Edwin Raja, J. Jeno Jasmine, P. Vimal Kumar, and R. Ramani. The future of well-being: Ai-powered health management with privacy at its core. In Bharat Bhushan, Akib Khanday, Khursheed Aurangzeb, Sudhir Kumar Sharma, and Parma Nand, editors, *Advances in Computational Intelligence and Communication Systems*, chapter 14. John Wiley & Sons, Hoboken, NJ, December 2024.

[2] Darrell L. Young, Perry Boyette, James Moreland Jr., and Jason Teske. Generative ai agile assistant. In Jessie B. Walker and Rebekah E. Jansen, editors, *Disruptive Technologies in Information Sciences VIII*, volume 13058, page 1305809. SPIE, International Society for Optics and Photonics, 2024. Presented at SPIE Defense + Commercial Sensing, National Harbor, Maryland, United States.

[3] Darrell L. Young, Eric C. Larson, and Mitchell A. Thornton. Prompt engineering for detecting phishing. In *Assurance and Security for AI-enabled Systems 2025*, volume 13476, page 1347609. SPIE, 2025.