**SecureRAG: Integrating Zero Trust Security into Dialogue-Based Human-AI Data Interactions**

The growing adoption of Retrieval-Augmented Generation (RAG) techniques with Large Language Models (LLMs) has transformed human-data interactions within organizations through natural language dialogues. While improving data accessibility, this evolution introduces socio-technical security risks, including data leakage and unauthorized information exfiltration during conversational interactions. Conventional security measures, such as static guardrails, are insufficient within dynamic environments where role-based access, needs to adapt to emergent situations and varying need-to-know requirements.  
  
To address these challenges, we present SecureRAG, a Zero-Trust security framework that integrates RAG methodologies with dynamic dialogue-based access control mechanisms guided by deontic logic. SecureRAG applies agentic, conversation-driven interactions, embedding security protocols as deontic rules (e.g., a belief about policies, permissions, obligations) enforced through user-agent dialogues. This approach enables active scrutiny, least-privilege access, and adaptive role-based permissions, aligning with Zero-Trust principles.

The SecureRAG is in the prototype stage and will begin productization in 2025 to support a physical security team stationed at an overseas allied partner nation installation, amidst strategic rivalry between great powers. The team, dispersed while patrolling, communicates through a secure group messaging app to quickly report anomalousevents, verify security protocols, and clarify *‘gray-zone’* activities. Given the area's local threat of espionage and data exfiltration, ensuring secure communication is critical.

Our solution embeds AI language model agents into the communication’s information retrieval chat framework. used for facilitating queries, routing data, and curating events. The goal is to create an ‘active understanding’ about what information is being queried from the knowledge base (which holds the event logs about threats, vulnerability assessments, and risk factors, to be synthesized for official reports). SecureRAG incorporates Zero-Trust principles, embedding security directly into conversations between users requesting information and the AI "curator" ensuring discussing and accessing sensitive data context remains protected through controlled dialogue negotiations during human-data interactions.

The solution methodology adapts to user queries and intent, providing proactive defense against insider threats and adversarial probing while preventing information exfiltration. By modeling security policies within interactive conversations, SecureRAG ensures that sensitive context and semantic insights are protected without impacting operational efficiency. Implementation of the approach looks to secure data, reduce cognitive load for users when accessing or logging security events, and improved situational awareness in dynamic environments with changing access levels and need-to-know restrictions.  
  
Our presentation will illustrate the system design for SecureRAG, including showing agent-mediated authentication and dynamic data access control via dialogue. This project work aligns with the "Cyber Architecture & Engineering (Protect)" capability pillar by integrating Zero-Trust principles into dialogue-based Human-AI systems for dynamic critical information protection.