Seminar Course Title: **AI Agentic Security**

Syllabus:

Foundations of AI Agentic Security: Definition of AI agents, threat models, comparison with traditional security.

Prompt Injection, Jailbreaks, and Semantic Firewalls: (i)Threats via natural language interfaces; (ii) Semantic firewalls: detection, enforcement, limitations.

Access Control for Agents and Secure Tool Use: (i) Capability-based security for tool invocation; (ii) Sandbox designs, API gateways, least privilege.

Data Security and Privacy: Training data poisoning, membership inference, leakage.

Adversarial Robustness for Agentic Systems: Multi-modal threats, certified robustness, verification.

Autonomy and Emergent Behaviors: Goal misgeneralization, long-horizon planning risks.

Trust, Provenance, and Identity of Agents: Authenticating agent actions, signatures, and provenance tracking.

Red Teaming and Incident Response for AI Agents: Vulnerability discovery, security playbooks.

Systemic and Societal Risks: (i) Multi-agent ecosystems, cyber-operations, misinformation; (ii) Governance, NIST AI RMF, EU AI Act.

Mitigations and Open Problems: Monitoring, auditing, interpretability, open research directions.