Adversarial Agentic AI (A3)

Infrastructure for Security Compliance Evaluation and risk Mitigation

# Problem Statement and Goals

Modern organizations operate in an increasingly hostile cyber environment, while facing stringent regulatory compliance obligations across sectors such as finance, healthcare, and cloud services. However, traditional cybersecurity defenses are often static, manual, and compliance verification is periodic, not continuous. This gap leaves organizations vulnerable between audits, and reactive rather than proactive in addressing vulnerabilities.  
Goals:  
We propose to design and build Adversa AI Agents Infrastructure, an autonomous agentic AI system capable of:

* Simulating Red/Blue/Purple team adversarial activities using AI on live or synthetic environments.
* Evaluating and strengthening an organization’s domain compliance posture dynamically.
* Mapping vulnerabilities and defenses directly to regulatory standards.
* Providing a continuous feedback loop for ongoing hardening of cybersecurity compliance.

# Description

Adversa AI is a modular agentic AI platform where intelligent Red Team and Blue Team agents autonomously engage in cybersecurity offense and defense across domain-specific environments. Red Team agents identify and exploit vulnerabilities, while Blue Team agents detect, respond, and reinforce system defenses.

A compliance mapping layer evaluates the impact of these interactions against regulatory standards such as ISO 27001, HIPAA, and PCI-DSS. The platform will be open-source at its core, supporting modular extensions for specific domains and compliance requirements.

# Approach: Three-Layered Modular Framework

The proect is structured into three primary layers:

* Core Open-Source Layer: Base Red/Blue agents and generic cyber environments.
* Domain-Specific Extension Layer: Adds context-specific attack/defense behavior (e.g., healthcare).
* Compliance Mapping Layer: Maps actions to compliance controls and scores adherence.

# Architecture

The architecture consists of a CyberArena environment (Docker-based), intelligent Red/Blue agents, domain-specific modules for behavior extension, and a compliance evaluation layer producing battle reports, compliance scores, and hardening recommendations. The Arena is built using the Microsoft Simulator Network open source software.

# Initial Focus

a) Web Hosting / SaaS: ISO 27001 and SOC 2

b) Healthcare Industry: HIPAA Security Rule

# Technical Elements

Base Model:  
Foundation LLMs like GPT-4 Turbo or Claude 3 will drive agent reasoning and planning. Tool use includes nmap, sqlmap, hydra.  
  
Prompt Engineering:  
- Role prompts, planning prompts (ReAct/CoT), feedback loops, and memory buffers.  
  
Infrastructure:  
- Docker-based simulation environment, logging services, and agent orchestration frameworks (LangChain, AutoGen).

# Development Plan

Phase 1: Build core environment and simple Red/Blue agent behaviors.  
Phase 2: Add compliance mapping to ISO 27001 controls.  
Phase 3: Expand with healthcare domain.  
Phase 4: Run evaluation and prepare outputs.

# System Simulation Strategy

Docker Compose-based cyber arena simulates:  
- Linux servers, web apps, EHR systems.  
- Configurable vulnerabilities (weak creds, open ports).  
- Future: Kubernetes and IoT simulation environments.

# Agent Performance Evaluation

Metrics:  
- Attack/Defense success rate  
- Reaction time  
- Compliance coverage score  
- Strategic improvement over time (agent learning)

# Outputs from the System

Key outputs:  
- Battle reports (timeline, actions)  
- Compliance evaluations (JSON, PDF)  
- Security scorecards  
- Suggested remediation steps