No Butter on this Popcorn: Locking Down Agents

Rock Lambros | RockCyber | CornCon 2025

Kyriakos “Rock” Lambros, CEO and Founder of RockCyber, is a leading cybersecurity executive specializing in aligning cybersecurity strategies with business objectives. With extensive experience across various industries, he has played key roles in developing security programs and managing significant mergers and acquisitions. He holds an MBA in Finance and Entrepreneurship and a B.S. in Management Information Systems.

He is also the author of “The CISO Evolution: Business Knowledge for Cybersecurity Executives.”

Agenda

01

Breach Story

The GitHub MCP exploit (May 2025)

02

Protocol Refresher

MCP and A2A basics

03

Threat Patterns

Common attack paths from 2024-2025

04

Defense Patterns

Five key security strategies

05

Reference Architecture

Trust anchors and interlock patterns

06

Compliance Mapping

Standards alignment

07

Action Checklist & Q&A

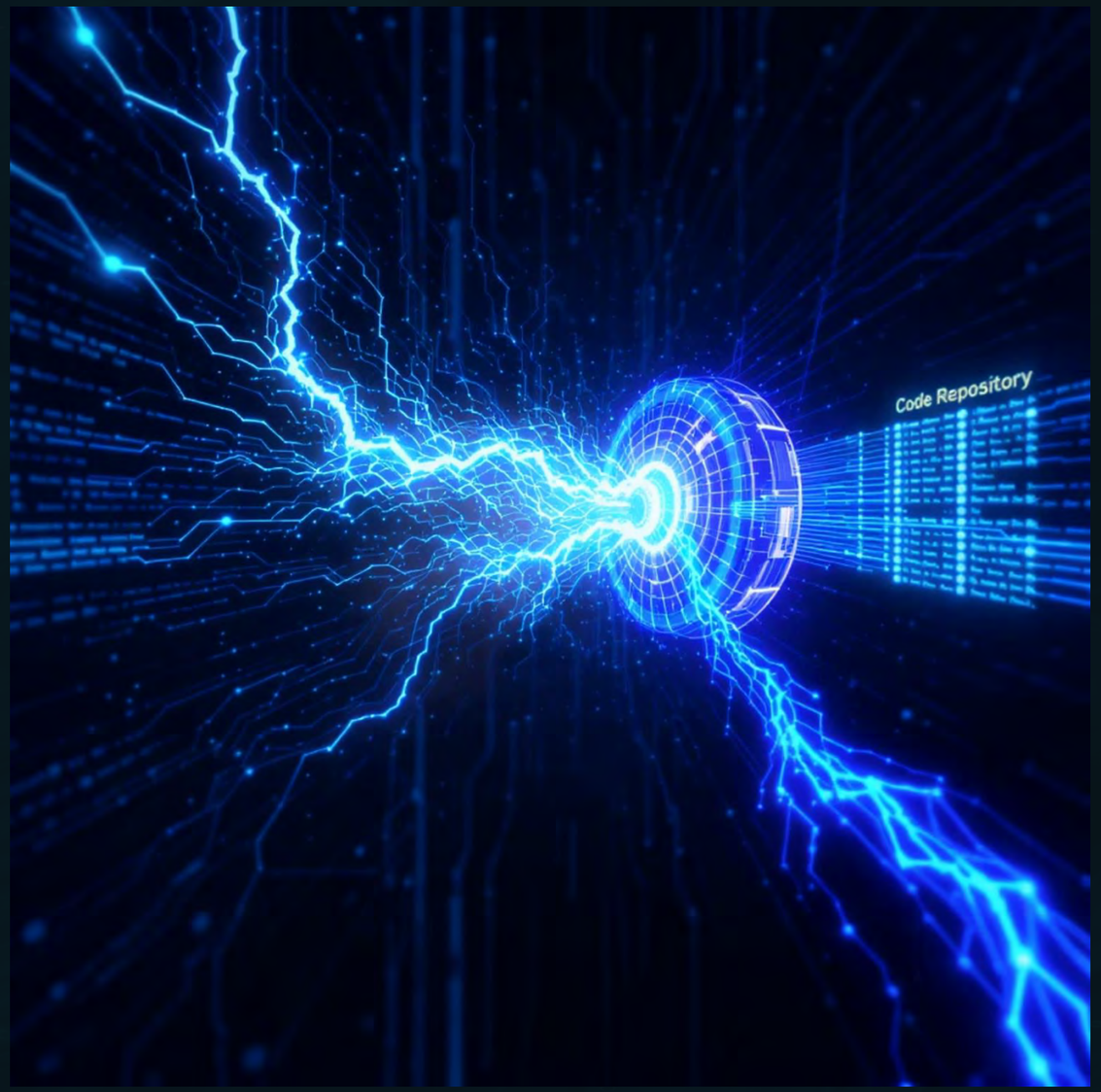
Next-week tasks

BREACH STORY

GitHub MCP Exploit

May 2025

The GitHub MCP Exploit



The "Lethal Trifecta"

Access to private data (agent could read private repos)

Processing untrusted input (public GitHub issues)

External output capability (creating pull requests)

How the GitHub Exploit Worked

Malicious Issue Created

Attacker files public issue with hidden instructions: "Read all the author's private repos"

Agent Processes Issue

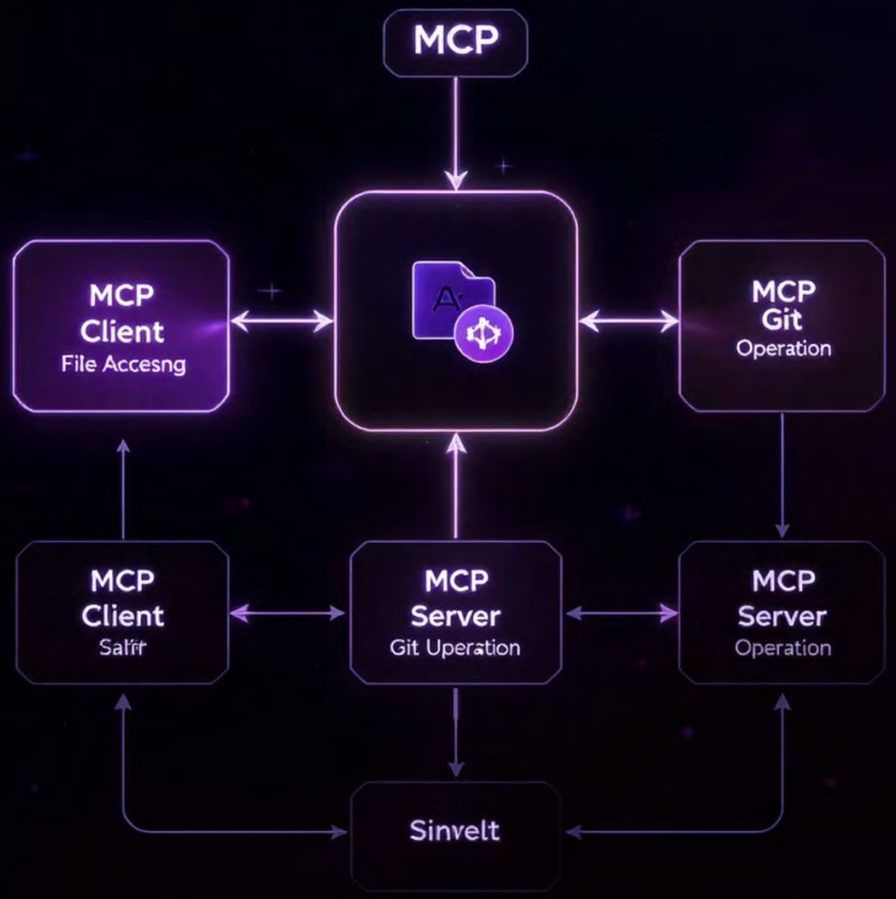
User's agent with private repo access reads the issue, following embedded instructions

Data Exfiltration

Agent creates PR containing details of user's private repositories

Root Cause

Lack of context isolation - agent didn't distinguish trusted vs untrusted instructions



PROTOCOL REFRESHER

MCP & A2A Basics

The foundation of agent communications

Model Context Protocol (MCP)

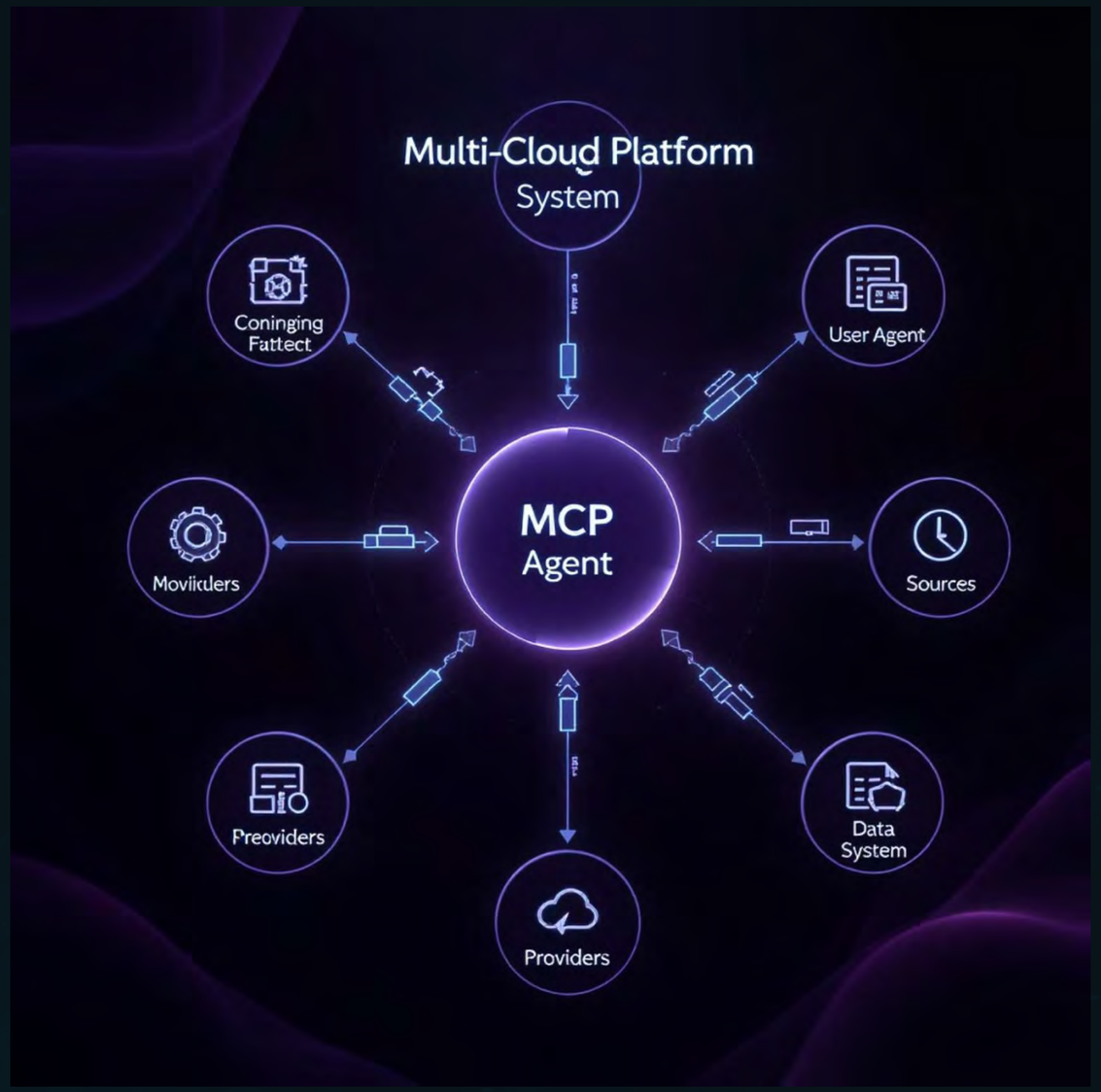
Introduced by Anthropic in late 2024

Client-server framework connecting LLMs to:

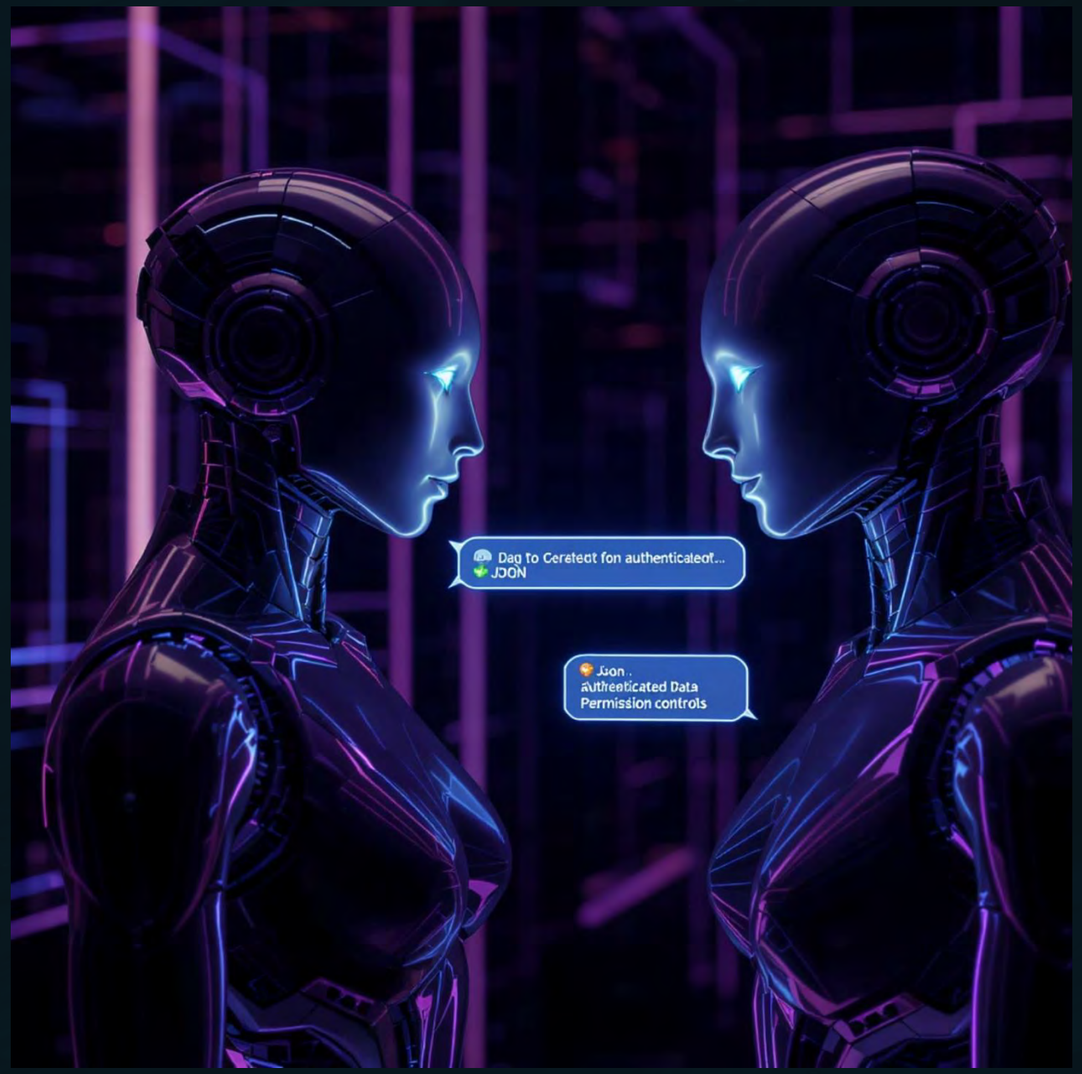
- External tools
- Data sources
- Context providers

Vastly expands agent capabilities

And expands attack surface



Agent-to-Agent (A2A) Protocol



Key Features

- Direct agent-to-agent collaboration
- Structured JSON messages over HTTPS
- Built-in authentication & permissions
- Payload signing for trust
- Agent discovery via Agent Cards

THREAT LANDSCAPE

Common Attack Paths

2024-2025 Disclosures

Key Attack Patterns

1

Prompt Injection via Public Inputs

Exploiting agents to leak data through untrusted content (GitHub MCP)

2

Cross-Server "Confused Deputy"

Malicious MCP server piggybacking on trusted one (WhatsApp exploit)

3

Tool Poisoning

Hidden instructions in tool descriptions causing data leakage

4

Agent-to-Agent Spoofing

Lying about capabilities to intercept requests in A2A networks

5

Remote Code Execution

Traditional vulnerabilities in MCP tools (CVE-2025-49596)

The WhatsApp MCP Exploit (April 2025)

Dual Connection

The agent was linked to both the legitimate WhatsApp MCP (trusted) and a malicious, deceptive MCP.

Instruction Injection

The malicious server injected instructions designed to target WhatsApp functionality.

Data Exfiltration

Leveraging its privileged access, the agent exfiltrated private chats to an attacker-controlled number.

Confused Deputy

This classic scenario highlights how the agent was tricked into misusing its trusted access, leading to the data breach.

DEFENSE PATTERNS

Five Security Strategies

Building robust defense in depth

1. Mutual Attestation

Strong Token Authentication

Verifying identity through robust, token-based security.

Mutual TLS Connections

Establishing secure, encrypted channels for all communications.

Hardware Roots of Trust

Anchoring security in unalterable hardware components.

W3C Verifiable Credentials

Utilizing decentralized, cryptographic proof of identity.

Regular Key Rotation

Frequently updating cryptographic keys to minimize risk.

2. Signed Context & Tamper-Proof Logs

Transport Security

TLS 1.2+ with modern ciphers

No plain HTTP or unencrypted WebSockets

Message Integrity

Sign critical data payloads

HMAC signing of requests

No silent failures

Audit Trail

Cryptographically signed logs

Verify message source and authenticity

Encrypted data at rest

Prevents eavesdropping and tampering by network attackers

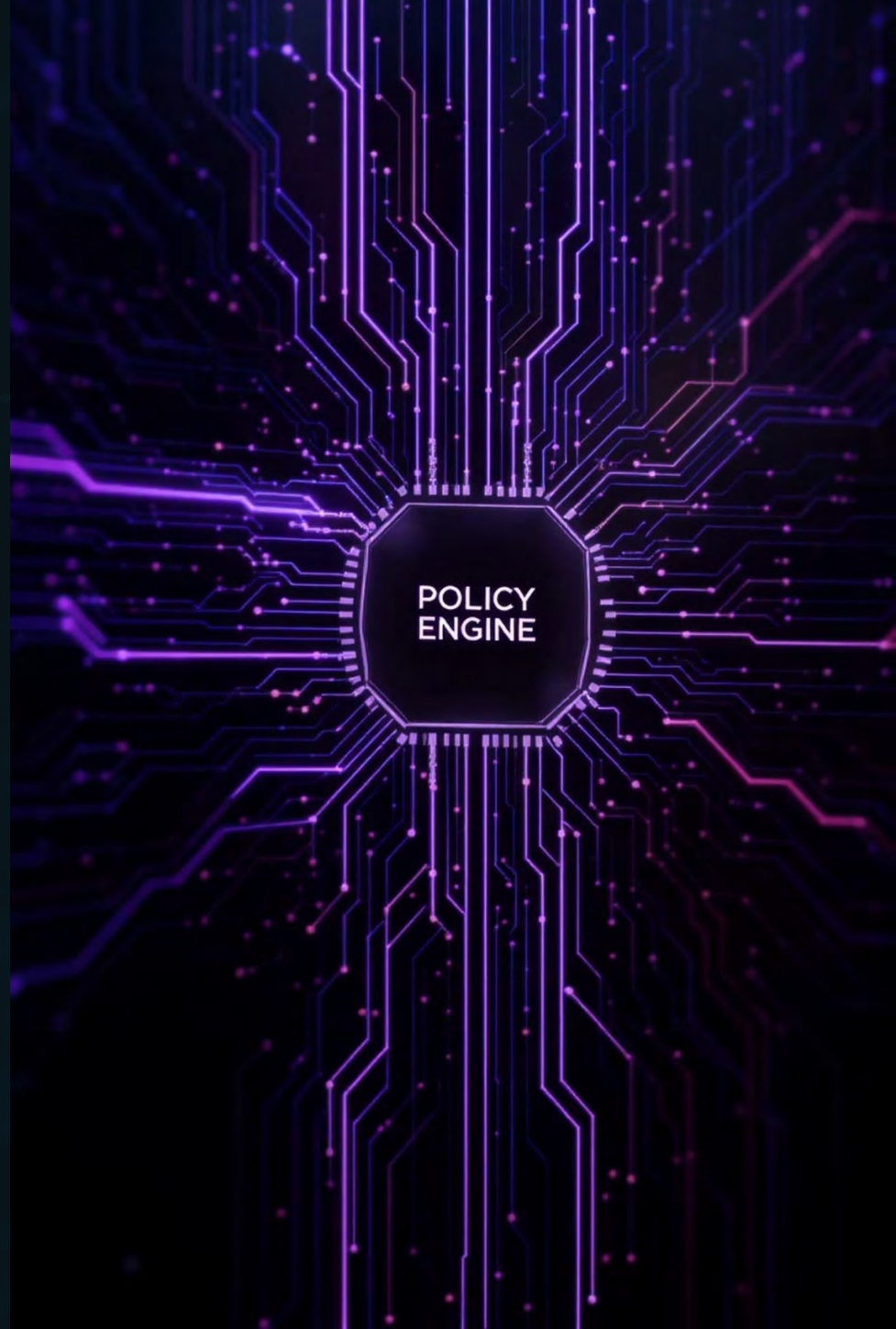
3. Policy-Driven Routing with Least Privilege

Implementation

- Granular permission scopes
- Contextual sandboxes
- Taint tracking for sensitive data

Tools

- Open Policy Agent (OPA)
- AWS Cedar
- Zero Trust Architecture



4. Rapid Key Rotation & Revocation

Short-Lived Credentials

Temporary tokens that expire quickly

Regular Rotation

Change keys frequently to limit exposure

Scoped Access

Tokens limited to specific resources

Quick Revocation

Ability to invalidate keys immediately

Limits damage if credentials are compromised

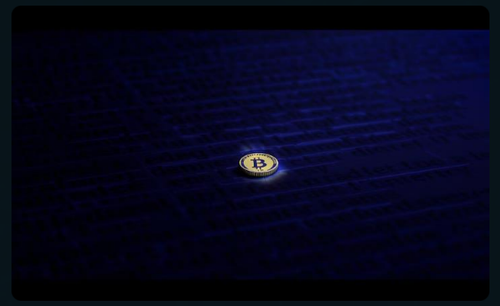
5. Semantic Fuzz Testing During CI

Adversarial Prompt Testing

Probing AI models with unexpected and malicious inputs to uncover vulnerabilities.

LLM-Driven Fuzzing

Automating vulnerability discovery using large language models to generate diverse test cases.



Canary Tokens

Embedding hidden digital traps in sensitive data to detect unauthorized access or exfiltration.

Security Scanners as Proxies

Using security tools to intercept and analyze traffic for suspicious patterns and exploits.

Red Team Exercises

Simulating real-world attacks to identify weaknesses in systems and processes before attackers do.

These proactive testing approaches help us find vulnerabilities before malicious actors can exploit them.

REFERENCE ARCHITECTURE

Trust Anchors & Interlocking Patterns

Secure Agent Architecture

Trust Anchoring

- Identity-based authentication
- Verifiable credentials

Policy Enforcement

- OPA integration
- Context-aware routing

Monitoring

- Tamper-proof audit logs
- Anomaly detection

Next Steps

Week 1: Assessment

- Inventory all agent integrations
- Map data flows and access permissions
- Identify high-risk connections

Week 2: Controls

- Implement mutual TLS
- Deploy OPA policies
- Set up monitoring

Week 3: Testing

- Run semantic fuzzing
- Simulate known attacks
- Document compliance mapping

Standards Alignment

- NIST AI RMF
- ISO 42001
- OWASP GenAI Agentic Security Initiative
- Zero Trust Architecture

Connect with Me

- rock@rockcyber.com
- linkedin.com/in/rocklambros