

KEYNOTE

# Defense in Depth for the AI Era

*Fighting AI-Enabled Attacks*

---

Rahul Patel • CIO • KBC USA

CISO New York - Corinium • September 2026

# Where We're Headed

1

## Setting the Stage: AI-Enabled Attacks

AI-detected vulnerabilities, automated exploitation, and new attack surfaces in models, prompts, and agentic infrastructure.

2

## Defense in Depth for the AI Era

A five-layer model — from asset foundations to AI-augmented detection, model security, and agent governance.

3

## The Path Forward

A practical maturity path and concrete next steps for your organization.



PART ONE

# Setting the Stage: AI-Enabled Attacks

Before we talk defense, we need to be honest about how far the offense has already come.

# AI-Detected Vulnerabilities (Used Offensively)

1

## AI/ML-powered fuzzing

Finds memory corruption, logic flaws, and edge cases far faster than manual testing.

2

## Automated code review by LLMs

Spots injection points, auth bypasses, and insecure deserialization at scale.

3

## Mining public repos & patch diffs

Attackers infer unpatched vulnerabilities from changelogs, shrinking the n-day window.

**The same speed advantage defenders want, attackers already have.**

# AI-Detected Exploits & Automated Attack Chains

1

## AI-assisted exploit generation

Turns a vulnerability description or PoC into a working exploit with minimal human tuning.  
Mean time to exploit has gone negative.

2

## Automated recon-to-exploitation pipelines

AI chains scanning, vulnerability matching, exploit selection, and initial access.

3

## Lowered skill barrier

Democratizes exploitation — less sophisticated actors can now run advanced attack chains.  
Time-to weaponize is compressing from weeks to hours in reported cases

**AI system can run the entire attack lifecycle (recon → exploitation → credential harvesting → exfiltration) with minimal human direction.**

# New Attack Surface: AI Models & Prompt Injection

1

## Prompt injection (direct & indirect)

Crafted inputs, or poisoned content the model ingests, manipulate model behavior.

2

## Jailbreaking & guardrail bypass

Circumvents safety and business-logic constraints built into the model.

3

## Data poisoning & model inversion

Corrupts training data or extracts sensitive information from a deployed model.

4

## Insecure output handling

Model output is trusted and executed downstream without validation.

**Treat every model the way you'd treat any other untrusted input surface.**

# New Attack Surface: MCP Servers & Agentic AI

1

## Agents as privileged actors

An LLM with tool-calling access (e.g., via MCP) is effectively a new kind of privileged identity.

2

## Indirect injection via untrusted content

A webpage, email, or document reaches an agent that has tool-calling permissions.

3

## Malicious or compromised connectors

Third-party MCP servers introduce fresh supply-chain risk.

4

## Excessive agency

Agents granted broader permissions than needed turn a small exploit into a large blast radius.

**The frontier attackers are probing right now — most threat models haven't caught up.**



**While AI compresses the attacker's timeline from discovery to exploitation, defense cannot rely on legacy response models. Organizations must pair core defense-in-depth architecture with AI-powered, autonomous controls capable of detecting and containing threats with uncompromising velocity and decisive precision.**

*This is the case for defense in depth, rebuilt for the AI era.*



## PART TWO

# Defense in Depth for the AI Era

Five layers — from foundational hygiene to autonomous-agent governance — built to counter AI-enabled attacks.

# The Defense-in-Depth Model, Reimagined



*New layers required: model layer, agent/tool layer, prompt & context layer — on top of everything you already run.*

# Know What You Have. Keep It Proven.

1

## Extended CMDB / Risk Register

Assets, business processes, impact, risk score, and ownership — mapped in one register.

2

## Continuous Control Monitoring

Real-time validation that controls are actually working, not just documented.

3

## Patch & Vulnerability Mgmt

Faster cycles to counter AI-accelerated exploit timelines.

4

## Segmentation & Zero Trust

Limits blast radius regardless of what caused the breach.

5

## Identity & Access Mgmt

MFA, least privilege, PAM — still the #1 control against automated attacks.

6

## SIEM & Logging

The visibility baseline every AI-driven tool is built on top of.

**Know what you have, prove your controls still hold — AI attacks exploit organizations that skipped this.**

# Matching Machine Speed with Machine Speed

1

## Behavioral & Anomaly Detection

Catches novel malware and polymorphic payloads that signature-based tools miss.

2

## AI-Powered SOC Copilots

Accelerate triage, reduce alert fatigue, auto-generate incident summaries.

3

## Automated Threat Hunting

ML models proactively surface weak signals across huge log volumes.

4

## AI-Managed Deception Tech

Honeypots and honeytokens that adapt to lure and study automated attackers.

5

## Continuous AI-Driven Pen Testing

Always-on validation replaces annual pen tests — testing defenses at the pace attackers now move.

**Human-only response times can't keep pace with AI-driven attacks.**

# Treat the Model Like Any Other Production System

1

## Input validation & sanitization

Treat all user and external input to a model as untrusted — the prompt-injection era's answer to SQL injection.

2

## Guardrails & output validation

Filter and verify model output before it's trusted or acted on downstream.

3

## Red-teaming & adversarial testing

Attack your own models before someone else does it for you.

4

## Access controls & data segregation

Rate limiting, authentication, abuse monitoring, and protected training pipelines.

**It needs its own security lifecycle — not an exemption from one.**

# Governing the Newest Privileged Identity in Your Environment

1

## Principle of least agency

Agents get the minimum tool access and permissions needed — nothing more.

2

## MCP server vetting & sandboxing

Treat third-party connectors like third-party code: review, isolate, monitor.

3

## Human-in-the-loop checkpoints

Required for high-impact actions — payments, data deletion, external communication.

4

## Full audit logging & kill switches

Every tool call traced; access revocable the moment behavior looks anomalous.

**The agent layer is the newest attack surface — it needs governance before it needs more capability.**

# Technology Alone Doesn't Close This Gap

1

## Updated security awareness training

Covers deepfake vishing, AI-written phishing, and indirect prompt injection via everyday documents.

2

## AI-aware incident response playbooks

Explicit scenarios for model and agent compromise, not just traditional breaches.

3

## Cross-functional AI governance

Security, legal, data science, and business functions aligned on one framework.

4

## Extended vendor risk & framework alignment

Assess AI model providers and MCP integrations against NIST AI RMF and OWASP LLM Top 10.

**Culture and process have to evolve at the same speed as the technology.**

# A Maturity Path, Not a Finish Line

1

## Foundational

Traditional controls solid: CMDB, CCM, patching, IAM, segmentation, and baseline monitoring in place.

2

## Adaptive

AI-augmented detection and response operational; model-layer controls and red-teaming established.

3

## Autonomous-Ready

Full agent/MCP governance, automated response, and continuous adversarial testing at production scale.

*Ask your team: where are we on this path today — and what's the one move that gets us to the next stage?*

## CLOSING THOUGHT

**AI is already in the attacker's hands.**

**The question is: is it in yours ?**

- 1 Inventory first: extend your CMDB to cover AI models and MCP connectors
- 2 Pilot one AI-augmented detection use case in the next 90 days
- 3 Apply least agency to every AI agent already in production

*Thank you | Questions & Discussion*