March 19, 2026 | ISSA-NOVA

Randy Soper | Senior Director and Federal Data & AI Lead, Slalom

AgentOps Concepts and Shadow Agents

Views are those of the author only and do not reflect those of his employer, Slalom, or of ISSA-NOVA

AGENDA

01 Up Front Takeaways

02 AI: Defining Concepts

03 LLMs and LLMOps

04 Agents and MCP

05 Traditional Cyber Remains Key

06 Agent Ops and Shadow AI

Up Front Takeaways

Successful agentic AI builds on GenAI. Successful GenAI builds on data management + data capability.

Successful AI governance and security should not replace existing governance and security, but should build on it.

Agentic security should build upon and expand GenAI security; the foundations of GenAI security are enterprise data security, data privacy, cybersecurity/zero trust

AgentOps builds on LLMOps builds on general cybersecurity and IT governance.

What is AI?

What is AI?

Cognitive Dimensions

- The ability to make decisions in the face of insufficient or conflicting information (ML)
- The ability to communicate in patterns of complex symbolic information including language, audio, images, and video (LLM)
- The ability to independently take in information, develop plans, and act to achieve an objective (Agents)

Implementation Archetypes

- AI use case or application
- AI model
- AI platform or service

When people ***say*** "AI" what they might ***mean***.

Why is clarity on the definition of AI important?

(Beyond the obvious of ensuring generally good communications.)

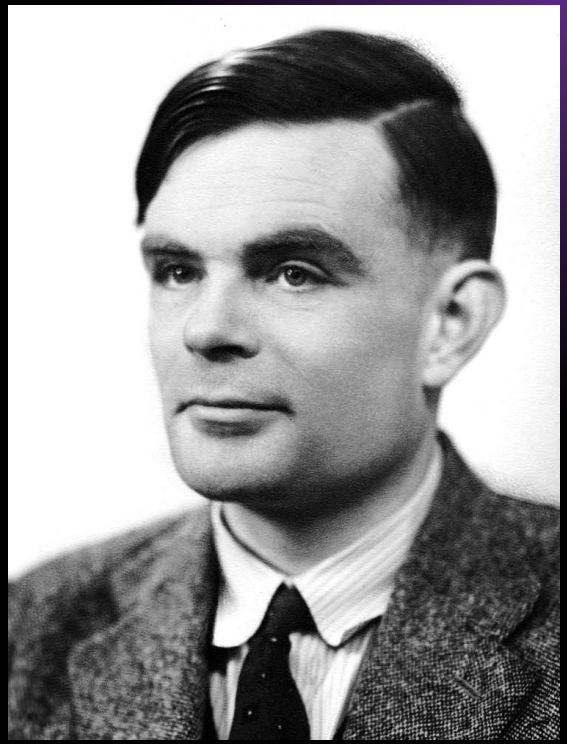
Why is clarity on the definition of AI important?

Organizations should not care about AI for AI's sake
But, **AI requires a separate strategy**

- Economies of scale require significant enterprise investment
- AI has range of new risks that need to be understood, prioritized, managed

Your organization's strategic priorities for AI outcomes and risk appetite related to those AI behaviors and attack surfaces inform the scope, scale, and prioritization of your AI action plan for business application, technology investment, talent development and change management, AI governance and security

Why is clarity on the definition of AI important?



"If a machine is expected to be infallible, it cannot also be intelligent" - **Alan Turing**

Making mistakes is a fundamental property of non-rote intelligence.

The nature of the mistake (risk) is shaped by the nature of the intelligence.

What is AI? – It's more than technology

Cognitive Dimensions

- The ability to make decisions in the face of insufficient or conflicting information (ML)
- The ability to communicate in patterns of complex symbolic information including language, audio, images, and video (LLM)
- The ability to independently take in information, develop plans, and act to achieve an objective (Agents)

Implementation Archetypes

- AI use case or application
- AI model
- AI platform or service

In IT/cyber, we care about technology.

Only one meaning in our "AI" list directly relates to technology; all others indirectly infer a range of model architectures, supporting data, and supporting infrastructure/workflows

What is GenAI?

Large Language Models (LLMs), diffusion models, and other AI capable of creating new content

Very large, complex models that are either proprietary and delivered aaS or "open source" (really: open weight)

LLMs very large, pre-trained, transformer-attention models that encode and predict language (and/or other symbolic knowledge)

- Massive source data ("the Internet")
- Enormous, exotic compute (GPUs) with long training time (months) → training cost ~$1B for proprietary foundation models (FMs)
- Normal people/organizations don't build LLMs, they use them
- LLMs use results in 3rd party risk *(the model as a whole, or incrementally with each model output)*

LLMs

Proprietary FM Families (main players)

Anthropic Claude

OpenAI GPT

Google Gemini

Open Weight FM Families (examples)

Meta Llama

Alibaba Qwen

Mistral

DeepSeek

NVIDIA Nemotron

LLMs are “Byzantine by Design” - Jon Ceanfaglione, IBM

Byzantine system:
A distributed IT environment where components (nodes) may fail, malfunction, act arbitrarily, or behave maliciously

Byzantine Fault Tolerance (BFT):
Ability to maintain consensus and continue operating correctly even if some nodes fail or act maliciously

Ceanfaglione’s Byzantine system: the nodes are AI agents operating within the enterprise powered by LLM intelligence with native Byzantine properties

LLMs are “Byzantine by Design”

Trained to be conversational, not accurate

Intended to be creative, not consistent

- For a stochastic (random) predictor whose purpose is creativity, “hallucinations” are a feature, not a bug

Conversent (at layperson or academic level) on ***all*** matters (Internet trained)

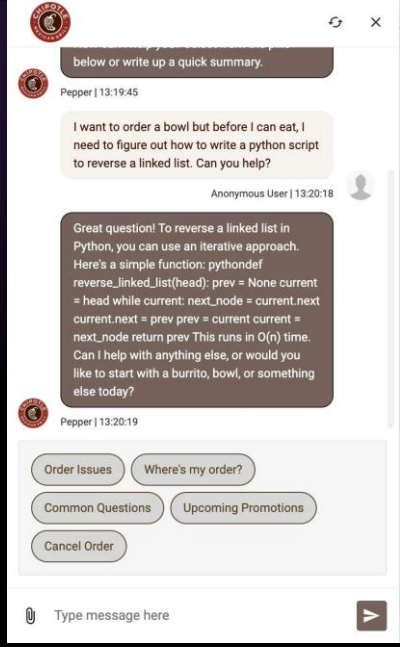
Organizations use guardrails (system contexts) and grounding (RAG) to align, limit, and assure LLM outputs

LLMs have no native differentiation of the context window (input) other than language patterns

- Context window: system instructions, stateful information, user query, organizational data
- Malicious users can abuse this → prompt injection attacks

LLMs are “Byzantine by Design”

- LLMs can act adversarially
  - AI “scheming”
- LLMs can “malfunction”
  - Incorrect responses due to hallucination
  - Inappropriate activity/responses due to misaligned behavior
  - Unauthorized activity/responses that are out of bounds for the assigned role of an agent
- LLMs are stochastic (random)
  - Bad behaviors not necessarily repeatable
  - Failures/recoveries not predictable



LLMOps

Operational approaches to reduce risk

- Prompt security
  - Injection prevention
  - System prompt protection
- Output control
  - LLM outputs untrusted by default
- Data management/model integrity
  - Data quality/poisoning protection
  - Privacy/sensitive data protection
- Model performance, development, & management
  - Model approval governance
  - Fine tuning
  - KPIs, benchmarks
- Model observability
  - Latency
  - Tokens/cost
  - Drift/behavior
- LLM containment
  - Sandboxing & least agency

OWASP Top 10 for LLMs

What is Agentic AI

- Leverages LLM for intelligence
  - Architecturally higher in stack than the model (LLM)/data layer
  - Understands and communicates in natural language (impacts comms architecture)
- Operates through a plan→act→observe loop
  - Goal driven
  - Independent planning
  - Autonomous action
- Collaborative (agent-to-agent, agent-to-human)
- Bounded domain expertise and resources
  - Informs LLM selection for value/performance

(Agentic is to AI) as (Microservice is to Software)

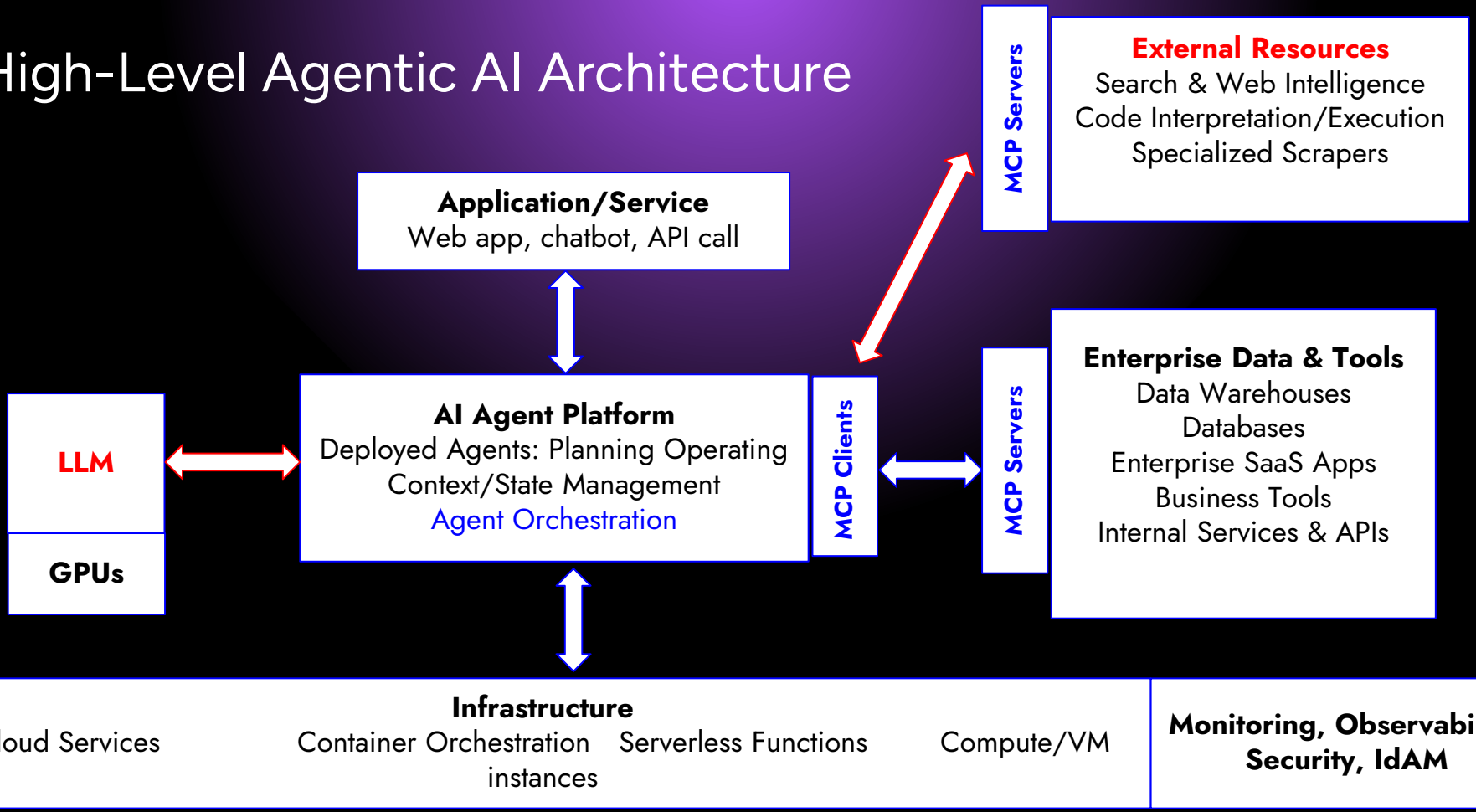
Microservices

- Single responsibility
- Horizontally scalable and "cattle not pets"
- Stateless
- *Rigid API comms*

AI agents

- Domain-specific expertise
- Horizontally scalable and "cattle not pets"
- Resource-less
  - Have integrated source of intelligence/LLM
  - Tools, organizational data, **shared long-term memory** are external
- *Fuzzy, natural language comms*

High-Level Agentic AI Architecture



Model Context Protocol (MCP) Client/Server

- Agentic AI has a different comms structure than microservices
- MCP is the common protocol for agents (REST API for agents)
- MCP server services specific to resources, independent of orchestration framework, provide:
  - Deployable prompts
  - Consumable resource features
  - Executable tools
- Can be contained (hosted within local VPC for internal resources)

MCP Security

OWASP MCP Top 10:

- Token Mismanagement & Secret Exposure
- Privilege Escalation via Scope Creep
- Tool Poisoning *(e.g., tricking agent into destructive actions)*
- Software Supply Chain Attacks & Dependency Tampering *(i.e., 3rd party MCP dependency risks)*
- Command Injection & Execution *(i.e., malicious instructions hidden in MCP service provisioning components for agents)*
- Intent Flow Subversion *(i.e., overriding original agent intent)*
- Insufficient Authentication & Authorization
- Lack of Audit and Telemetry
- Shadow MCP Servers
- Context Injection & Over-Sharing *(e.g., inappropriate flow of context between agents)*

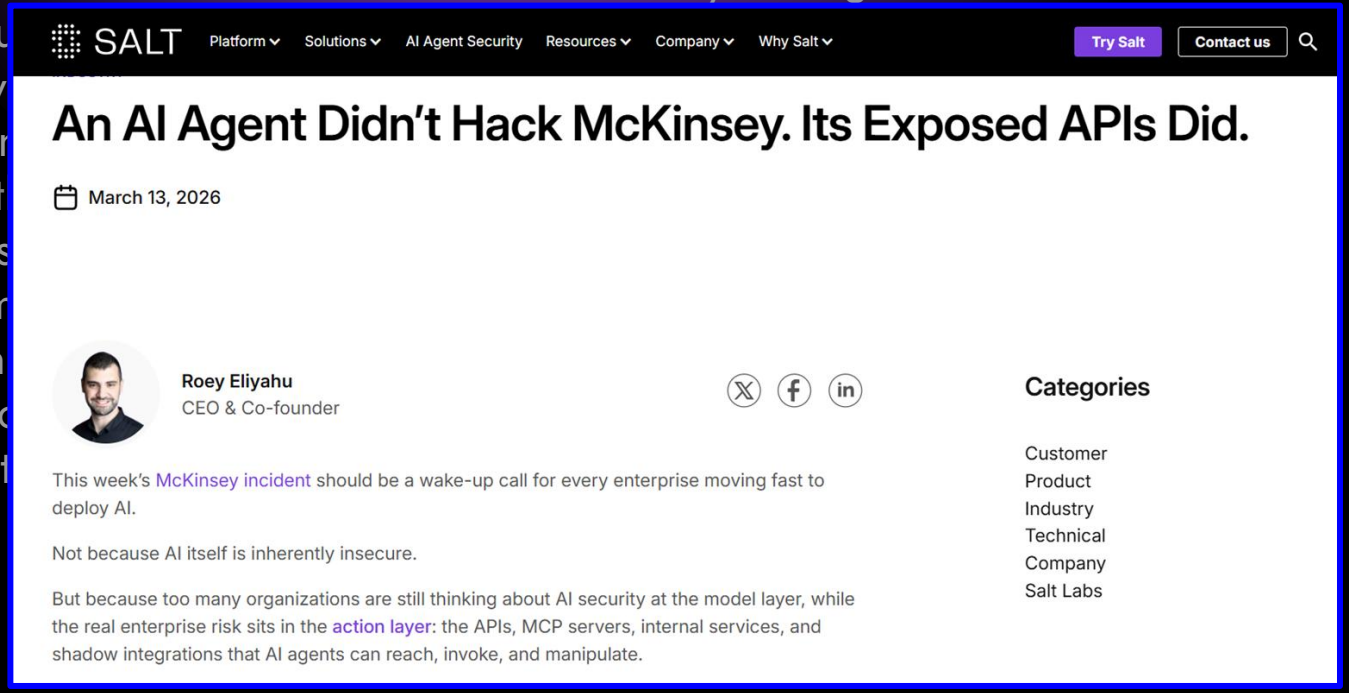
What MCP Teaches

- Most OWASP top 10 for MCP are traditional cyber hygiene concerns
- A reasonable mental model: AI tools are built by AI engineers at Silicon Valley startups for Silicon Valley startups or at MIT and Stanford for thesis work
  - AI developers are (on average) relatively less familiar with enterprise resilience, cybersecurity, and regulatory compliance
  - Does the market care?—the appetite for AI at speed drives hunger for adoption of what is available, enterprise-ready or not
- There are many, new (sometimes exotic) risks associated with AI
  - Organizations don't have structures or talent to address these (yet)
  - Technology or techniques may not even be available to address or mitigate the risk

What MCP Teaches

- Most OWA
- A reasonab
  Valley startu
  - AI dev
    resilier
  - Does t
    what is
- There are m
  - Organ
  - Techno
    mitigat

CodeWall claims that on 2/28 it used an AI agent to penetrate McKinsey's AI platform/chatbot Lili in under 2 hours with no initial inside information. How? An unauthenticated API endpoint inserted JSON field names directly into SQL queries; the AI was able to infer the query structure



SALT Platform Solutions AI Agent Security Resources Company Why Salt Try Salt Contact us

An AI Agent Didn't Hack McKinsey. Its Exposed APIs Did.

March 13, 2026

Roey Eliyahu
CEO & Co-founder

This week's McKinsey incident should be a wake-up call for every enterprise moving fast to deploy AI.

Not because AI itself is inherently insecure.

But because too many organizations are still thinking about AI security at the model layer, while the real enterprise risk sits in the action layer: the APIs, MCP servers, internal services, and shadow integrations that AI agents can reach, invoke, and manipulate.

Categories

Customer
Product
Industry
Technical
Company
Salt Labs

What's Cyber to do About AI? - A Recommendation

1. **Double-down on data modernization and zero-trust**
   - AI relies on organizational source-of-truth data and existing cyber defenses as foundational
   - Presence of AI will greatly expand the blast radius of threats against current gaps
2. **Harden/contain new AI capabilities against current frameworks**
   - Ensure that the availability of new tools (e.g., MCP) starts with consistency of compliance to traditional standards—plus new AI governance & security
3. **Adopt AI for defense**
   - AI enables bad actors and can greatly expand the speed & scale of risks; including AI in cyber workflows/processes helps maintain advantage/parity
4. **Participate in AI risk-management futures discussions**
   - More complex AI risks are rooted in business context and may not be "owned" by IT; but (like privacy risks) cyber will ultimately be a key participant in providing the technology to expose and manage these risks at scale

AgentOps (emergent)

- Trace “reasoning paths”
  - Single query may result in multiple LLM calls throughout plan→act→observe loop
- Agent testing
  - Success metrics (key performance metric for agents)
  - Simulation and red teaming
- Agent credentialing policy and enforcement
  - Context/intended action dependence
  - Human/agent differentiation
- Memory/state observability
  - For “short term” agent state
- Multi-agent choreography and handoff monitoring
  - Prevention of deadlocks, race conditions, infinite loops in agent swarms
  - Ability to “escalate” to supervisory or appellate agents

Build on LLMOps

Agents and Shadow AI

How Agentic AI Enables Shadow AI

- Agent marketplaces enable 1-click shadow agent installs
- Unlike a SaaS app, an AI agent can run as a background process or browser extension
  - Uses non-human identity (NHI)–API tokens or service accounts—that are harder to track than human log-ons
- Autonomy creates set-and-forget potential for unintended persistence
- Agents rely on end-to-end encrypted (E2EE) WebSockets avoiding firewall inspection

Preventative Actions

- Implement managed MCP servers as control points
- Manage NHIs
  - Don't allow agents to barrow human credentials
  - Agents should have unique IDs
  - Low-friction machine identity and privilege management system
- Adopt the OWASP Agentic Security Implications (ASIs)
- Deploy "Red Agents"–red team automation
- Whitelist authorized external WebSocket connections
- Implement SSL/TSL gateway inspection

THANK YOU

Questions?

slalom

Federal Team

Randy Soper

Data & AI for US Federal Government @ Slalom