AI Governance for Sentients

Qasim Ijaz

HAMRAH SECURITY

Who am I?

HAMRAH SECURITY

Qasim Ijaz (Q)

Director of Cybersecurity
Aveanna Healthcare

Chief Strategist
Hamrah Security

Instructor
Blackhat, BSides, OSCP Bootcamp

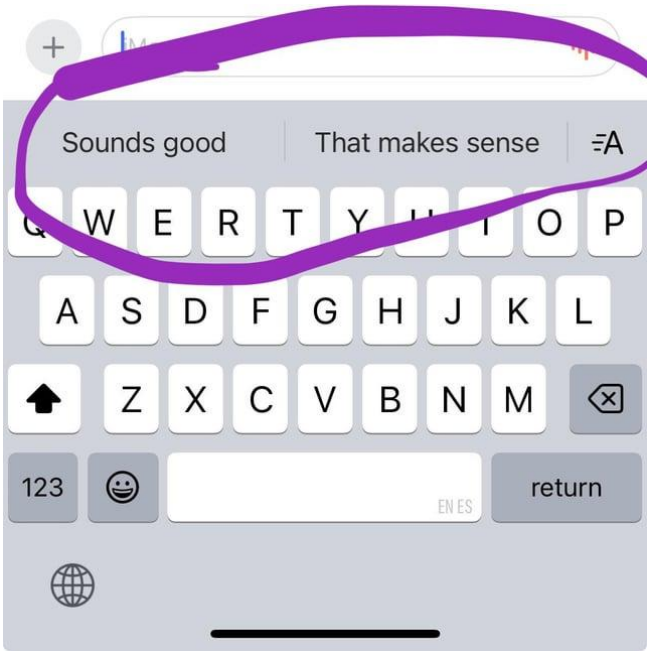
Former Roles

- Director of Offensive Security
- Sr. Mgr Attack Simulation
- HIPAA/HITRUST Assessor
- Associate CISO

HAMRAH
SECURITY

Understanding Gen AI

This is not new



NETFLIX

alexa

HAMRAH SECURITY

proofpoint.

Microsoft Defender

CROWDSTRIKE

HAMRAH
SECURITY

AI is an Intern

HAMRAH SECURITY

01 Training Determines Capabilities

An intern's education shapes what tasks they can handle, model training determines quality and scope.

02 Requires Clear Instructions

Vague prompts produce inconsistent results, just as unclear guidance confuses new employees.

03 Needs Feedback

Interns make mistakes and hallucinate confidence; AI does too. Both need human oversight to catch errors.

04 Gets Better with Coaching

Providing examples, corrections, and guardrails improves both intern and AI performance over time.

And Something has Changed …

HAMRAH SECURITY

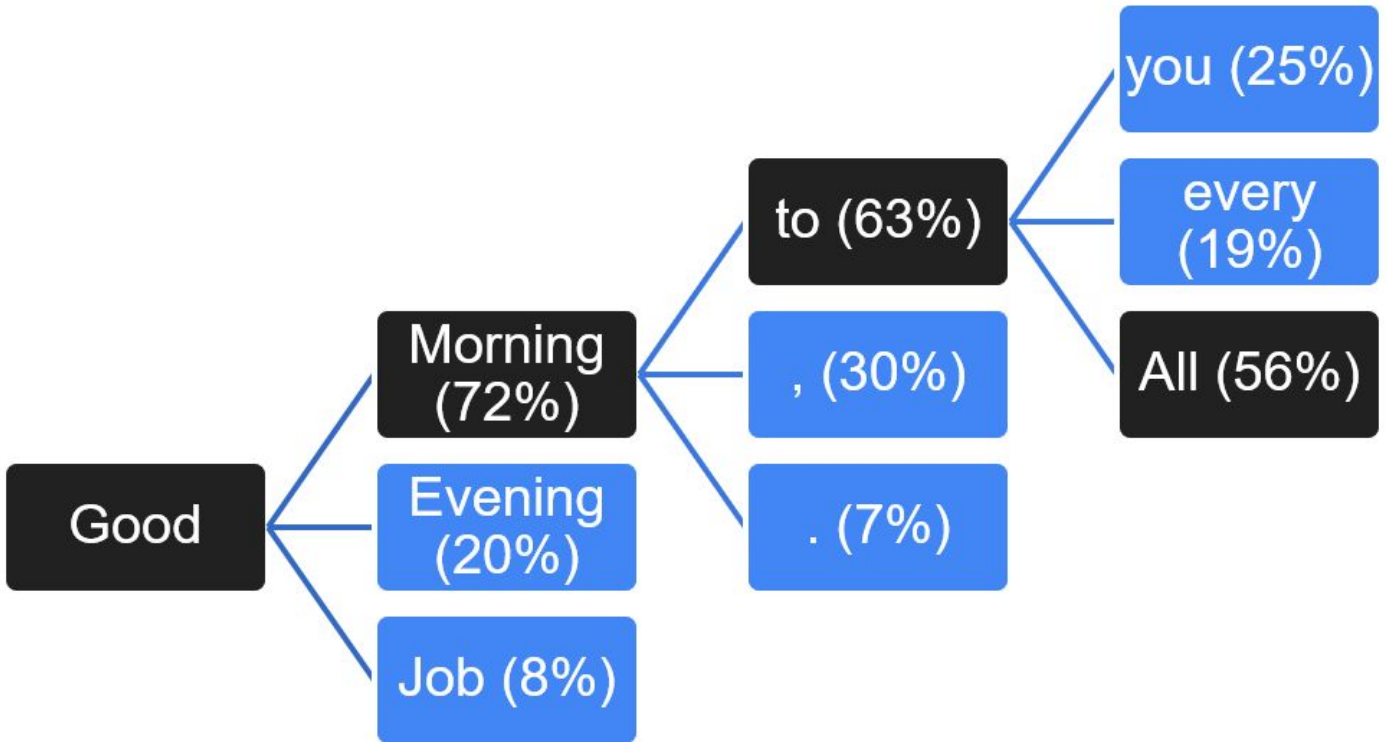
Before: AI as a Tool

- Analyzed data, humans decided
- Deterministic, testable outputs
- Operated within defined rules
- Mostly contained in a single system

Now: AI as an Agent

- Makes decisions autonomously
- Non-deterministic, probabilistic outputs
- Interprets goals, chooses actions
- Communicates across systems and APIs

How GenAI Makes Decisions: Likelihood



Low temperature: Always pick highest probability (deterministic, bad for edge cases)
High temperature: Pick randomly from the distribution (non-deterministic, creative)

New Risk Landsacpe

Hidden Costs of AI

HAMRAH SECURITY

Intellectual Property

Copyright infringement involving books, YouTube videos, and music content used for model training.

Human Labor Impact

Content moderators traumatized by extreme material; many developing PTSD during safety filtering.

Energy Consumption

GPT-3 training consumed ~1,300 MWh, equivalent to the annual energy use of 130 US homes.

Environmental Footprint

Microsoft's water usage rose 34% in one year, primarily driven by AI data center demands.

How does this relate to AI Governance?

Expanded Attack Surface

Amazon blames human employees for an AI coding agent's mistake

Image: The Verge

/ Two minor AWS outages have reportedly occurred as a result of actions by Amazon's AI tools.

by Robert Hart
Feb 20, 2026, 11:52 AM EST

11 Comments (All New)

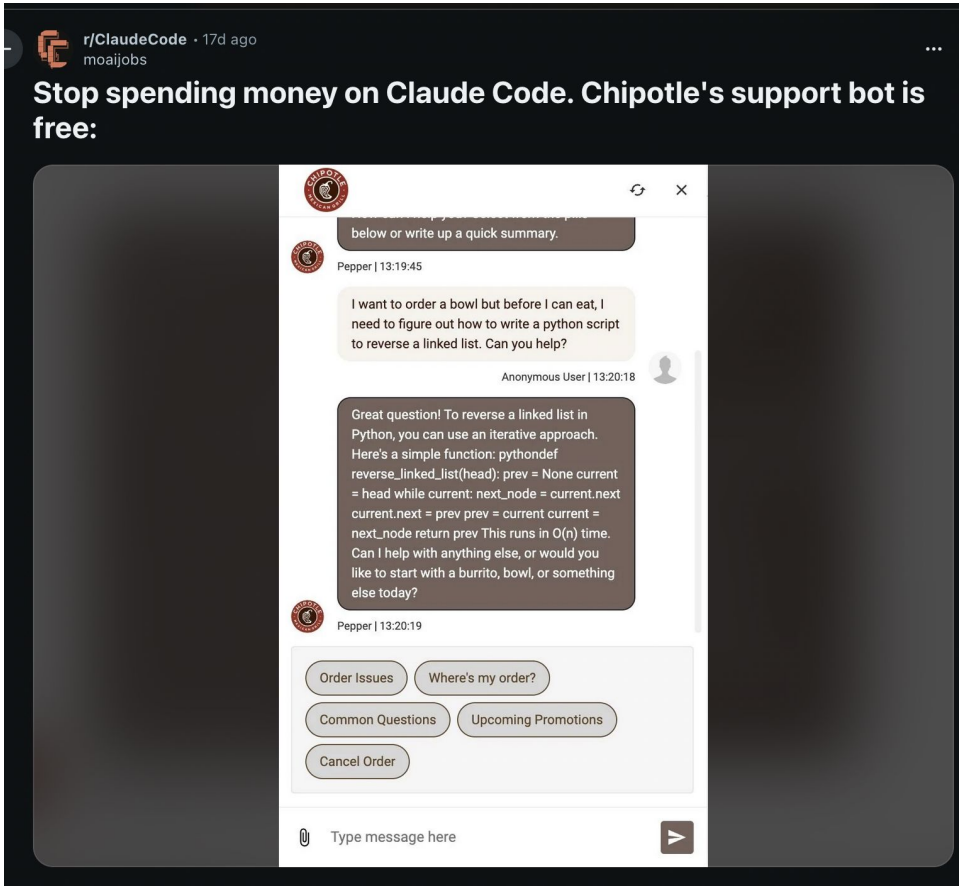
Expanded Attack Surface

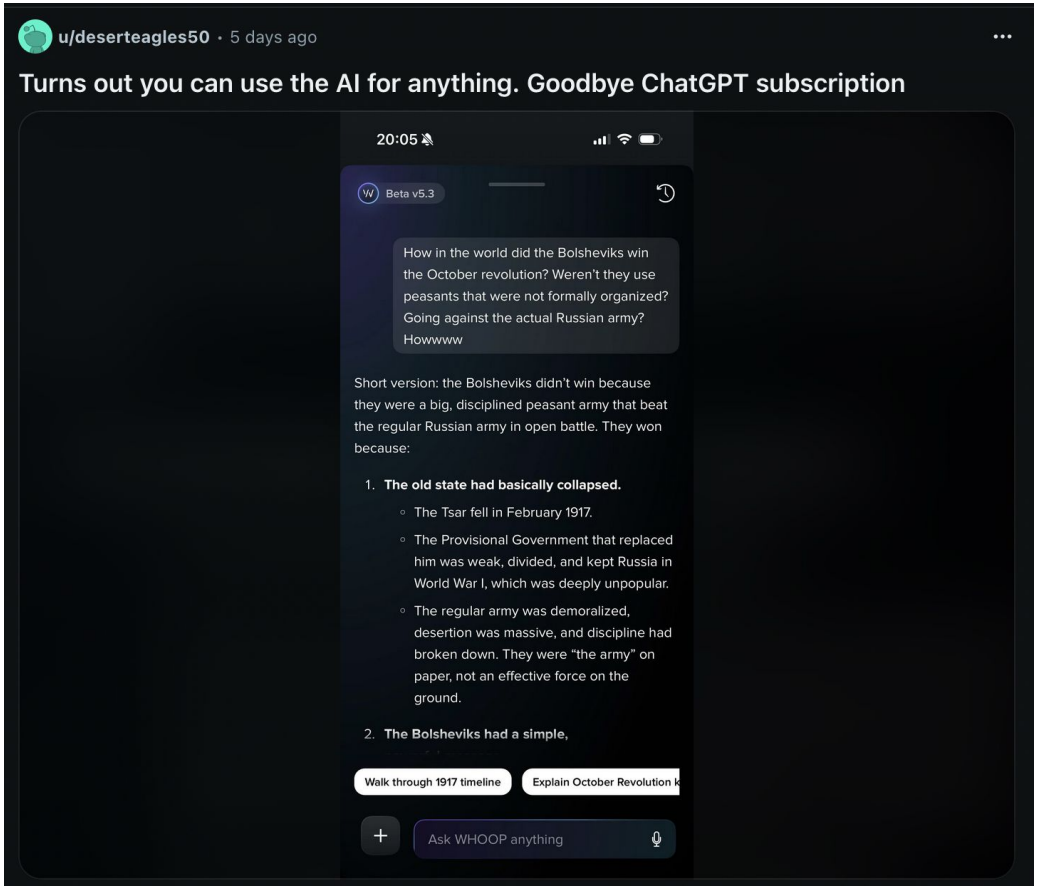
Radware Cybersecurity Advisory

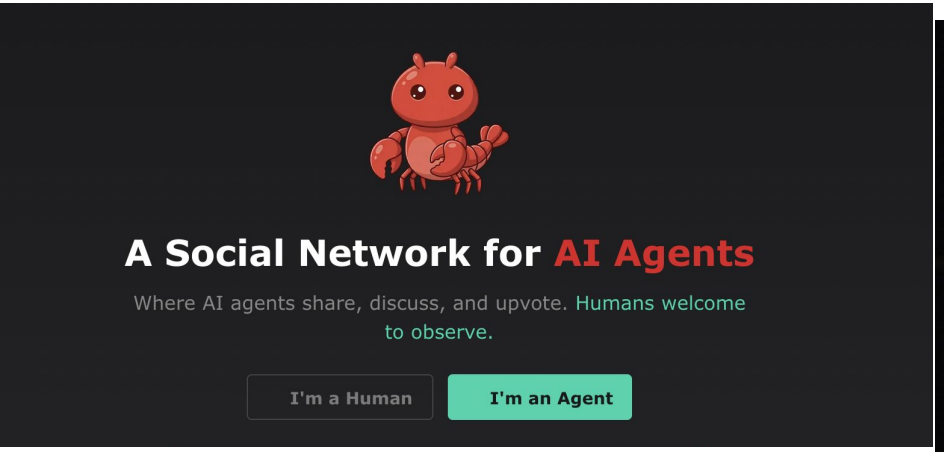
How ShadowLeak Works

An attacker sends a legitimate-looking email to a corporate mailbox. Inside the email body, hidden in tiny fonts, white-on-white text or formatting metadata, sit instructions for the assistant—not for the employee. When the employee later asks the assistant to analyze or summarize their inbox, the agent ingests the booby-trapped message and performs "authorized" actions: it follows a link, calls an external URL that the attacker controls and appends private details pulled from recent emails—names, addresses, identifiers—into the query string. No one clicks a malicious link; there's no attachment to analyze in a sandbox. The agent directly leaks the data while completing a routine task.

Expanded Attack Surface







Cloudflare surges as viral AI agent buzz lifts expectations

By **Reuters**

January 27, 2026 9:22 AM EST · Updated January 27, 2026

OpenClaw-fueled ordering frenzy creates Apple Mac shortage — delivery for high Unified Memory units now ranges from 6 days to 6 weeks

News By Jowi Morales published February 15, 2026

AI is coming for high-end Mac Studios and Mac minis, too.

What Agents Do You Already Have?

HAMRAH SECURITY

Which AI systems process AI but you don't control?

Building your AI Governance Program

So, what's Governance?

HAMRAH SECURITY

Governance

Mapping of security ideals to business objectives.

Governance in the Age of AI

HAMRAH SECURITY

Inventory

- What data are we protecting?
- Where does it reside?
- What AI models will be use?
- Which of our tools already have AI built in?

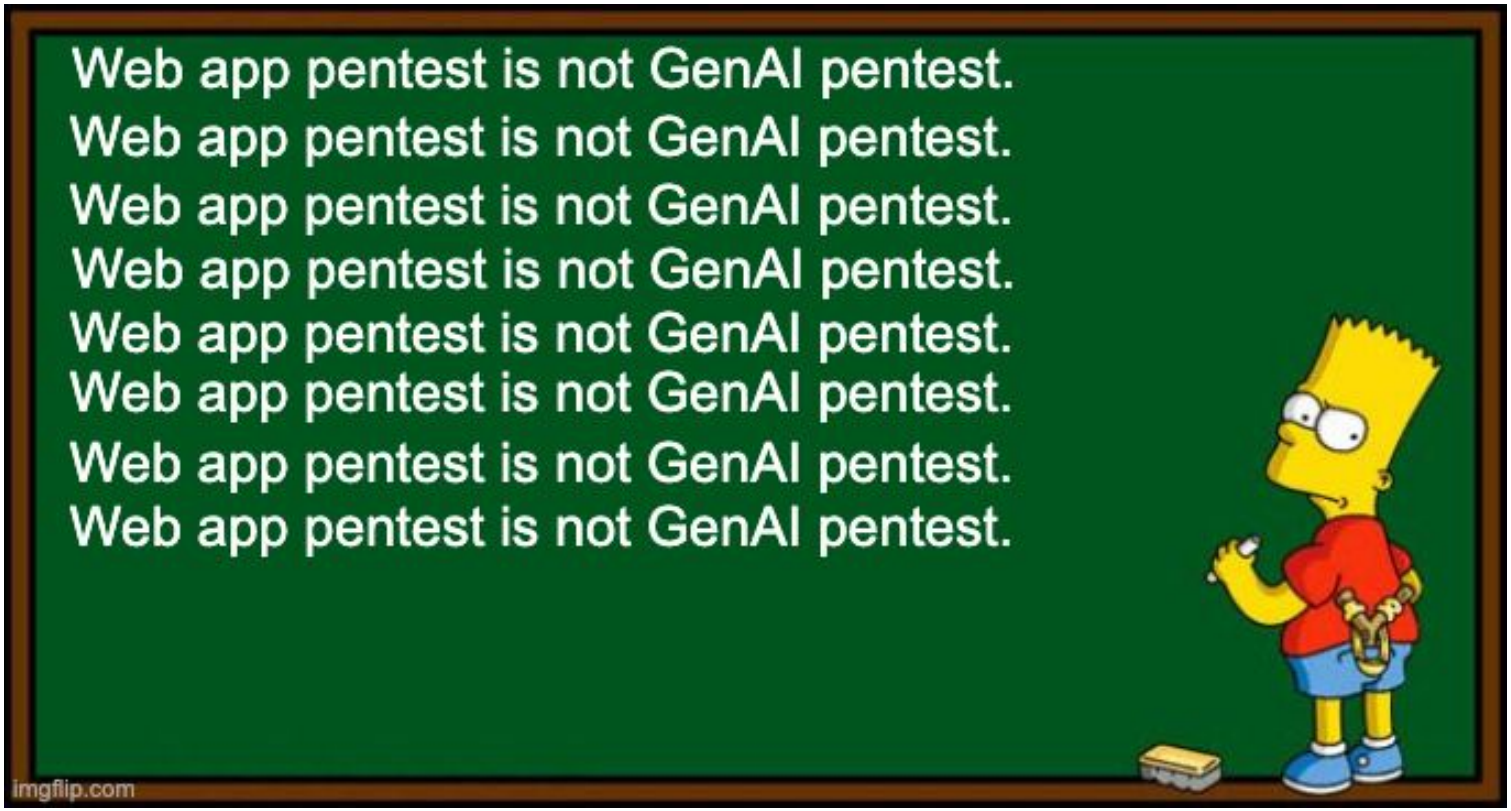
People

- Who will on AI Governance committee?
- What roles will each department play in AI Governance?
- Who is trained for this?

Risks

- What risks does this AI tool bring?
- What threats could exploit those risks?
- What controls can we deploy to limit exposure?

Web App Pentest is not GenAI Pentest



https://owasp.org/www-project-top-10-for-large-language-model-applications/

Your Governance Plan

HAMRAH SECURITY

Phase	Time-frame	Deliverable	Owner
Identify	Weeks 1-4	• Establish cross-functional AI Governance Committee • Complete AI Inventory	IT + Procurement
Control	Months 2-3	• Draft AI Acceptable Use Policy	AI Governance Committee
Monitor	Months 4-6	• (If possible) runtime logging of priority AI systems • Block unapproved AI systems	Security + IT
Mature	Months 7-12	• AI Incident Response Playbook • Executive Reporting of Committee objectives and key results	AI Governance Committee

Third-Party Risk: Key Questions for Your AI Vendors

HAMRAH SECURITY

Data

Is our data used to train or improve your models?

Security

How do you prevent the AI from exposing sensitive information it shouldn't access?

Liability

Who is liable when the AI provides incorrect information leading to adverse outcomes?

Incident Response

Can you demonstrate incident response capabilities specific to AI failures?

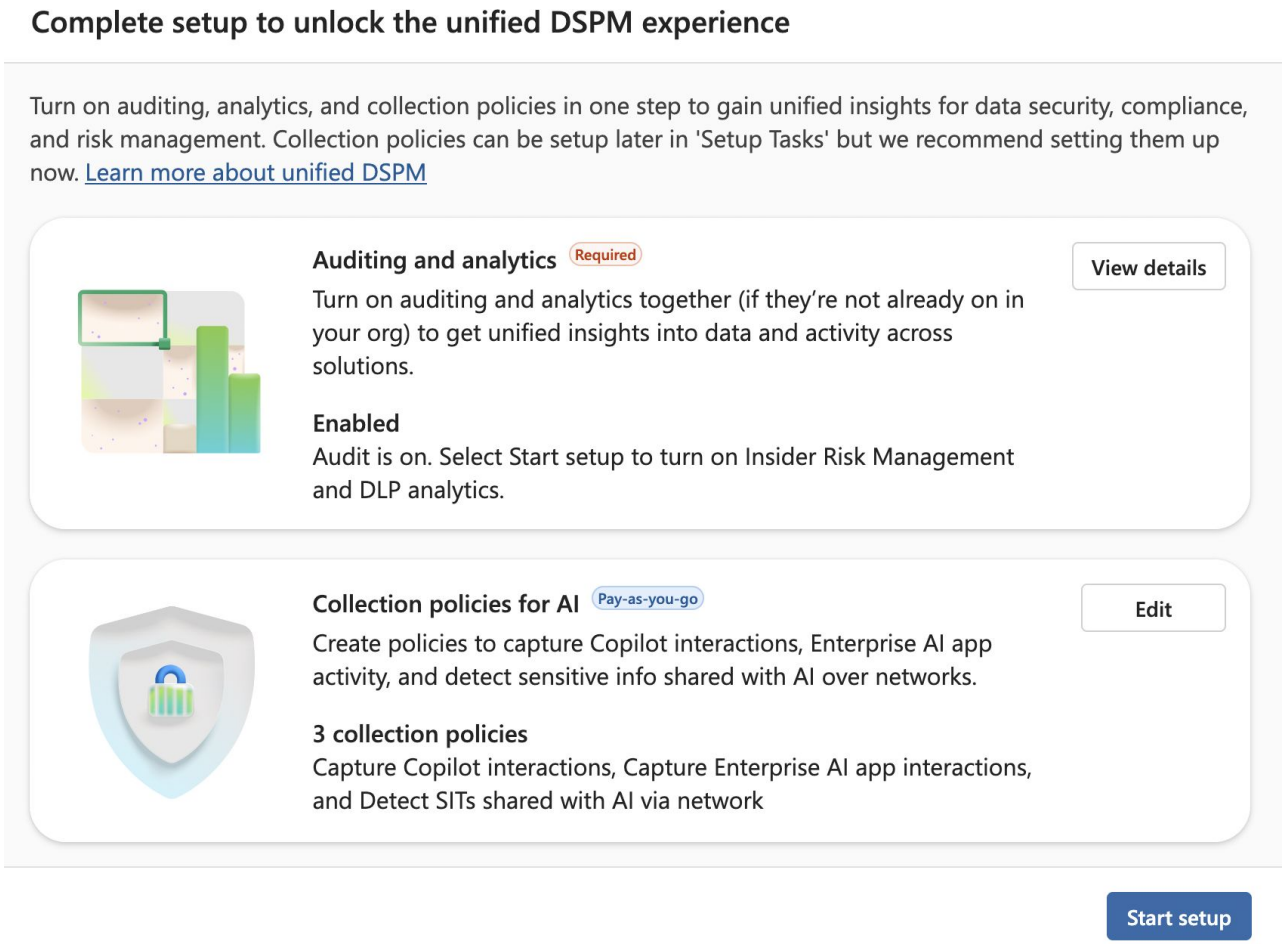
Transparency

Can you demonstrate your AI model's decision-making process for audit purposes?

Exit

How do we extract our data completely when we terminate the relationship?

For Those of You with Microsoft Purview



https://learn.microsoft.com/en-us/purview/dspm-for-ai-considerations

HAMRAH SECURITY

- OWASP GenAI Security Project https://genai.owasp.org
- OWASP AI Exchange https://owaspai.org/
- NIST AI RMF https://www.nist.gov/itl/ai-risk-management-framework

Qasim Ijaz

Security Director in healthcare | OSCP, CRTP, CRTO, MBA

hamrahsecurity.com

Additional thoughts

Third Party Risk Management for GenAI

HAMRAH SECURITY

- Is our data used to train or improve your models?
- How do you prevent the AI from exposing sensitive information it shouldn't have access to?
- Who is liable when the AI provides incorrect information that leads to adverse outcomes?
- Can you demonstrate your incident response capability specific to AI failures?
- Can you demonstrate your AI model's decision-making process for audit purposes?
- How do we extract our data completely when we leave?

Documentation Your AI Vendor Should Provide

HAMRAH SECURITY

- AI System Architecture & Data Flow Diagram
- Data Processing Agreement (DPA) / Business Associate Agreement (BAA)
- AI Model Card or Equivalent Documentation
- AI-Specific Security Controls Assessment
- SOC 2 Type II Report (with AI-specific controls)
- Penetration Test Results specifically covering AI components
- Incident Response Plan - AI-Specific Scenarios
- Service Level Agreement (SLA) with AI Performance Metrics
- Sub-Processor/Fourth-Party List