Booz Allen

McDermott Will & Schulte

When AI Breaks Bad:

The New Playbook for Incident Response

Privacy + Security Forum
Fall Academy
Nov 12-14, 2025 | Washington, DC
George Washington University, 800 21st St NW

Disclaimer: The content of this document is for informational purposes only. Booz Allen Hamilton is not responsible for your reliance on, or implementation of the guidance described in this document.

boozallen.com

Agenda

- AI Threat Landscape Overview
- AI Threats & Attacks
- Use of AI by Threat Actors
- AI Systems Attack Surface
- Security Considerations of AI Systems
- Legal and Regulatory Landscape
- Incident Response Considerations
- The AI Incident Response Playbook
- Preparing for AI Incidents
- Final Word

Presenters

Alex Southwell

Partner
McDermott Will & Schulte

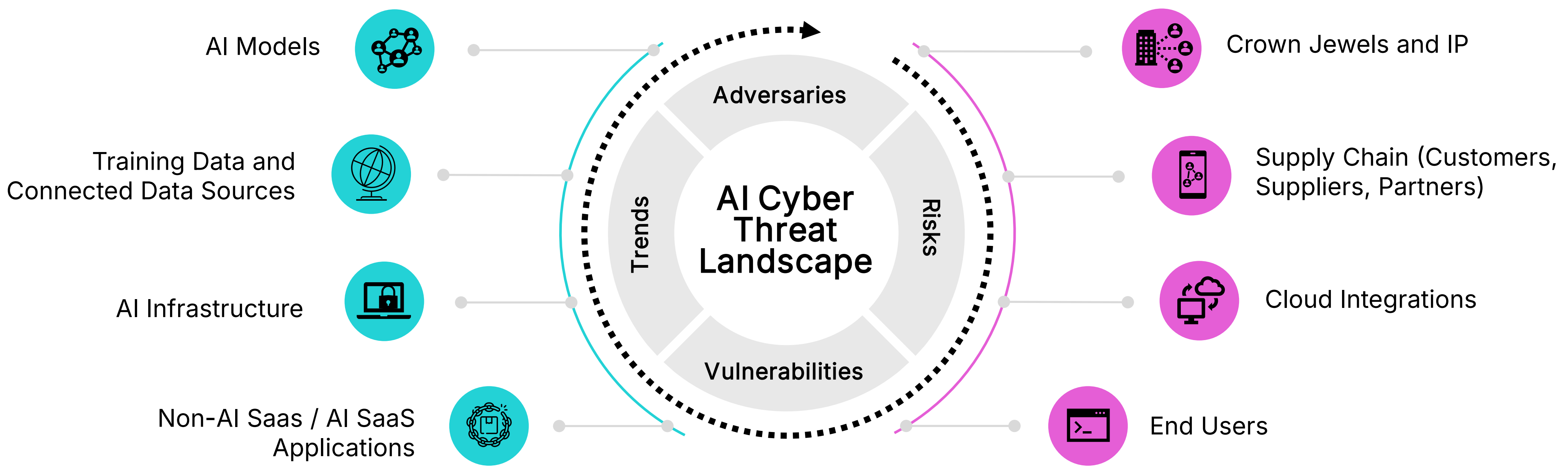
Brendan Rooney

Senior Vice President
Booz Allen Hamilton

Tony Gaidhane

Vice President
Booz Allen Hamilton

The Dynamic Threat Landscape | A(I) Growing Problem



Direct AI Threat Targets

- AI Models
- Training Data and Connected Data Sources
- AI Infrastructure
- Non-AI Saas / AI SaaS Applications

Indirect AI Threat Targets

- Crown Jewels and IP
- Supply Chain (Customers, Suppliers, Partners)
- Cloud Integrations
- End Users

Common Threat Actors

- *Nation State*
- *Criminal*
- *Hybrid*

Targeted Attacks on:

- AI Models
- Training Data and Connected Data Sources
- AI Infrastructure
- SaaS Applications (AI and non-AI)

Leading to Impacts on:

- Crown Jewels and IP
- Supply Chain ecosystem
- Cloud Integrations
- End Users

Diverse Threats and Attacks across AI Ecosystems:

↗ **Data and Model Poisoning**: Adversaries insert malicious data into training or inference pipelines, subtly corrupting models. A 2024 MITRE ATLAS study showed that poisoning just 0.1% of training data could cause targeted model misbehavior.

↗ **Prompt Injection and Adversarial Input Manipulation**: Attackers embed hidden prompts (often in HTML comments or code snippets) that redirect or manipulate model outputs—e.g., a website embedding hidden instructions for a crawler-based LLM.

↗ **Credential and API Key Theft**: OAuth and API tokens are now highly targeted, as seen in the 2025 Salesloft/Drift breach where token compromise enabled access to Salesforce-integrated AI automations.

↗ **Model Extraction and IP Theft**: Systematic querying of hosted APIs (such as OpenAI or Hugging Face endpoints) can allow reconstruction of proprietary models.

↗ **Supply Chain and SaaS Vulnerabilities** : Compromised libraries, models, or MLOps tools can insert malicious dependencies into training workflows, possibly impacting customers, partners and suppliers.

↗ **Guardrail Bypass and Jailbreaking**: Attackers exploit prompt engineering to override safety systems (“DAN,” “developer mode,” or “ignore instructions”) to exfiltrate data or generate harmful content.

↗ **Infrastructure and Cloud Abuse**: Compromised GPUs or cloud compute instances are repurposed for crypto-mining or data exfiltration.

Observed Use of AI by Threat Actors

Use of AI in multiple areas

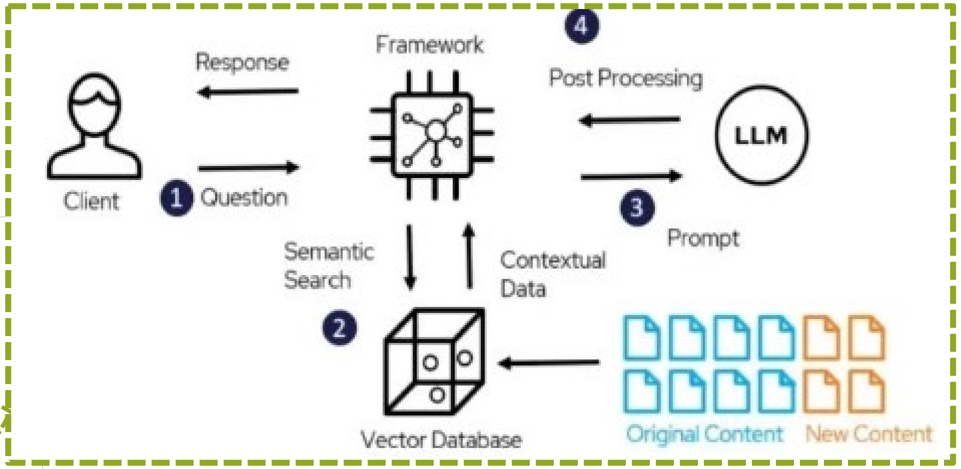
- **Content Generation:** Use of AI tools to generate emails, videos, voice or combinations, as well as malicious or regular code
- **Research and Efficiency:** Use of AI to research and exploit vulnerabilities, automate workflows, contextualize the attack
- **Malicious AI Model:** Use of AI Models to enhance tools, techniques and procedures (TTPs)

AI in Use by Threat Actors

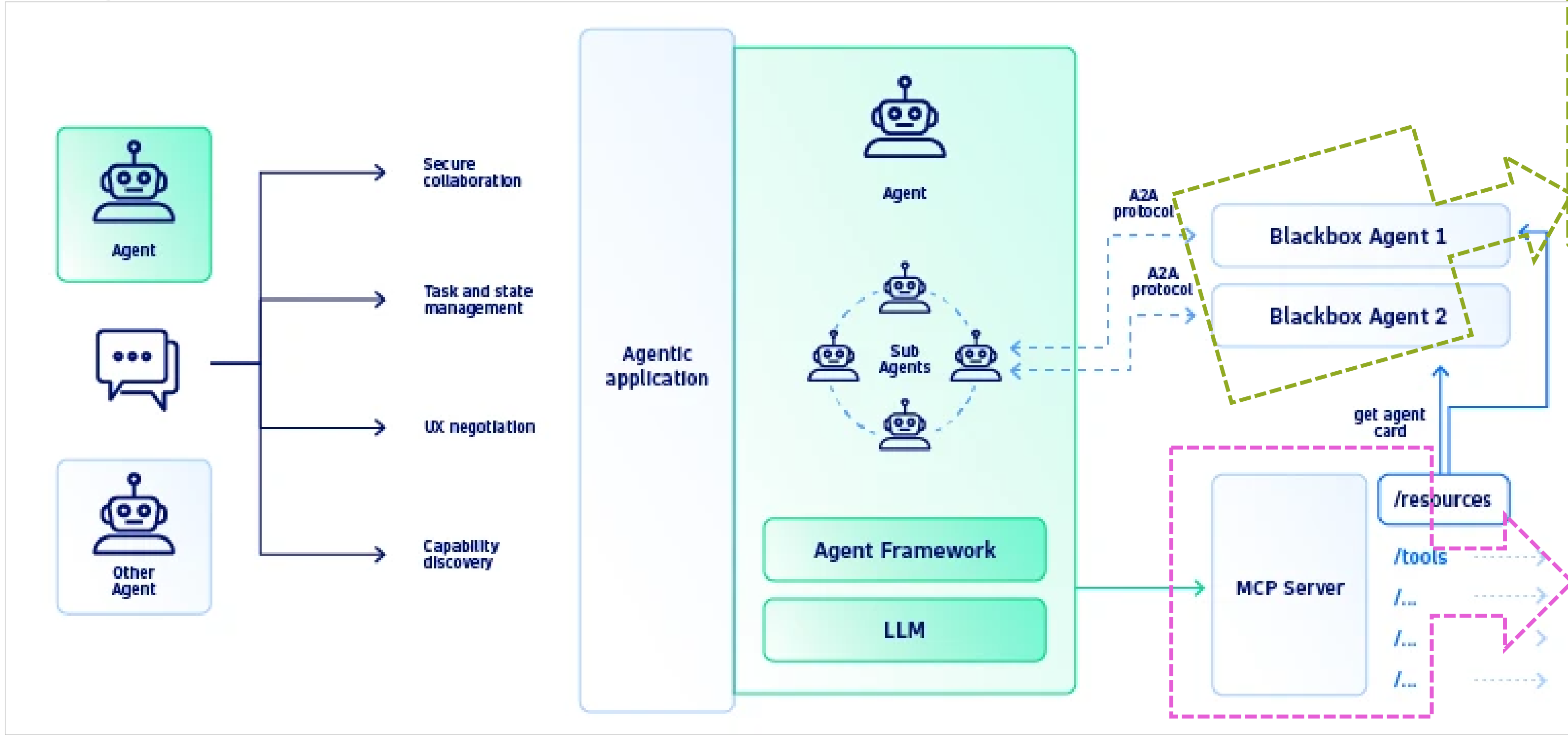
- **AI-Assisted Malware:** Adversaries now use LLMs and generative tools to create and obfuscate malware. Google's 2025 GTIG report identified new malware families (e.g. PROMPTFLUX, PROMPTSTEAL) that query LLMs at runtime to generate malicious scripts instead of hard-coding them.
- **Enhanced Phishing & Disinformation:** Attackers use AI to craft believable lures, spear-phishing messages, and deepfake content. AI-generated text and voices make social engineering campaigns more effective.
- **Code Generation & Reconnaissance:** Criminals leverage AI (ChatGPT, Codex, etc.) to write exploit code, gather system information, or analyze vulnerabilities. For instance, Iranian threat group MuddyWater used Gemini to research custom malware, improving their payloads.
- **State-Sponsored AI Models:** Notably, APT28 (Fancy Bear) deployed an AI-driven backdoor ("LameHug"/PROMPTSTEAL) in its C2 infrastructure. In mid-2025, LameHug used an open-source LLM to generate file-exfiltration commands on the fly.
- **AI Toolkits & Markets:** The underground economy now includes AI-powered attack kits. Google observed a maturing marketplace (2025) selling multi-purpose AI tools for phishing, malware dev, and vulnerability research.

Exponential Attack Surface of AI Systems

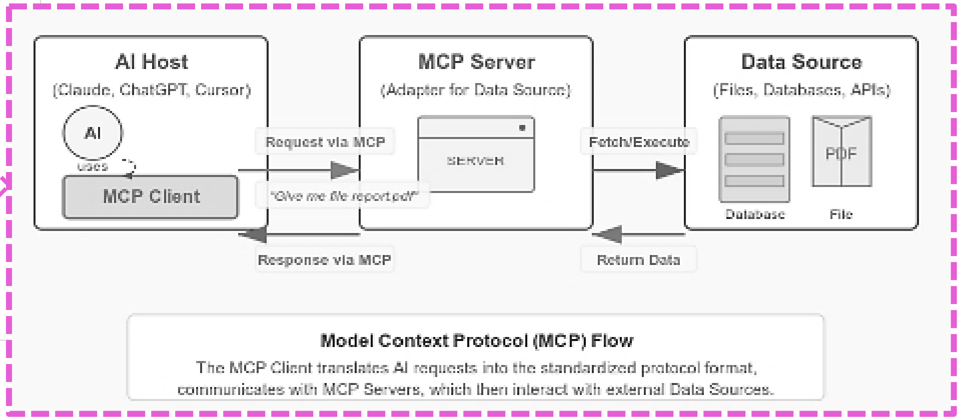
RAG System Architecture:



Agentic AI Based System:



MCP System Architecture:



Model Context Protocol (MCP) Flow

The MCP Client translates AI requests into the standardized protocol format, communicates with MCP Servers, which then interact with external Data Sources.

Image Sources:

- *RAG Model by deepchecks.com*
- *Agentic AI by Dynatrace*
- *MCP Graphic by Diamantia.substack*

Security Considerations of AI Systems

Defending AI systems requires protecting multiple layers: data, models, infrastructure, and usage

- ↗ **Training and input data must be secured** (e.g. data provenance, filtering out malicious inputs) to prevent poisoning or leakage
- ↗ **Trained model artifacts** (weights, configurations) should be treated as sensitive assets: use encryption, strict access control, and cryptographic signing so tampering is detectable. Infrastructure (cloud containers, servers) hosting AI must be hardened with network segmentation and runtime monitoring; inference APIs should require strong authentication and rate-limiting
- ↗ **AI supply chain** : vet third-party models and libraries (maintain a model SBOM) to avoid embedded vulnerabilities
- ↗ At runtime, **guardrails and monitoring** are critical – collect telemetry for every AI interaction (user prompts, model responses, vector store queries, etc.) and watch for anomalies like model drift or unusual query patterns.
- ↗ **Governance and compliance** (e.g. documented AI policies, risk assessments) ensure the organization stays ahead of evolving threats. In summary, key defense areas are: securing data and training pipelines, protecting model assets, hardening deployment environments, managing access/credentials, and enabling logging and monitoring tailored to AI systems.

Layer	Risk	Mitigation Strategy
Data	Poisoned, biased, or leaked training data	Vet data provenance, use checksum validation, differential privacy
Model	Tampering, extraction, drift	Cryptographic model signing, integrity verification, version control
Infrastructure	API & GPU exploitation, lateral movement	Network segmentation, workload isolation, runtime protection
Access & Identity	Token theft, insider misuse	Short-lived tokens, zero-trust authentication, centralized secret management
Observability	Lack of telemetry	Enable model logging (prompts, outputs, RAG queries), monitor for anomalies

Legislative Themes Relating to AI

- Transparency
- Consumer protection
- Employee protection
- No Bias / discrimination
- No election interference
- Privacy protection
- No deepfakes

AI Laws and Regulations

No comprehensive federal AI law

State Laws:

- **Colorado Artificial Intelligence Act**
  - First broad regulation of "high-risk AI systems“
  - Applies to developers and deployers, with important exceptions, including financial
  - Imposes reasonable duty of care to avoid "algorithmic discrimination“
  - Effective Jun. 30, 2026

- **Texas Responsible Artificial Intelligence Governance Act**

- **New York Responsible AI Safety and Education Act (RAISE Act), S6953B/A6453B**
  - Pending action from Gov. Hochul
  - Applicability driven by spending, not revenue
  - Primarily a transparency bill – mandates safety testing and incident reporting

- **European Union– EU AI Act (Regulation (EU) 2024/1689)**
  - First comprehensive AI legislative framework globally
  - Classifies AI systems by risk level, from minimal to unacceptable (e.g., social scoring)
  - Depending on risk levels, requires transparency, conformity assessments/registration, and human oversight
  - Applies to a wide array of services, including financial services
  - According to the European Securities and Market Authority (ESMA):
    "While AI holds promise in enhancing investment strategies and client services, it also presents inherent risks, including algorithmic bias, data quality issues, and (potential) lack of transparency."
  - Entered into force Aug. 2024, with the EU Commission planning phased implementation, enforcement, and assessments from 2024-2030

Why AI Incidents are hard:

- Multiple **interdependent components** (LLMs, RAG, APIs)
- Incomplete **telemetry** (prompts, model versions, embeddings)
- **Autonomous agent** actions complicate containment
- **Poisoned data** persists beyond traditional rollback
- **Legal & IP implications** of model tampering
- Ripple effects to **connected systems and data**
- Unforeseen **Recovery and Containment** implications for AI Systems

AI incidents differ drastically from conventional cyber incidents due to multi-layered complexity:

- **Stack Diversity:** AI systems combine LLMs, RAG modules, model orchestration (e.g., Model-Context-Protocol), SaaS integrations, APIs, and underlying cloud infrastructure (compute, networks, storage). This diversity means attacks can occur at any layer. For example, an adversary might poison the RAG knowledge base (context poisoning), corrupt long-term memory, or inject hidden prompts via a user input – all without ever breaching traditional servers.
- **Limited Logging:** Most AI APIs don't capture user prompts or model reasoning states, hindering forensic analysis. Typical SIEM logs don't capture LLM prompt/response histories or vector DB queries by default. Moreover, SaaS-based AI tools may provide limited telemetry.
- **Data Dependencies:** Poisoned vector databases or training data can persist across retraining cycles.
- **Dynamic Code Paths:** LLM agents may autonomously call APIs, execute code, or query databases, complicating containment.
- **Investigation**: Incident handlers must track prompt logs, model version hashes, API usage, and vector-retrieval logs, often aggregating from disparate sources.
- **Containment is non-trivial**: isolating a live AI service may require revoking API tokens and rerouting traffic, while "clearing" a compromised model might mean rolling back to a prior version.

The AI Playbook for Incident Response

AI IR Focus Areas:

- ↗ **Forensics:** Refresh skills, knowledge and tooling
- ↗ **Restore** clean model checkpoints or **retrain** if needed
- ↗ Rotate **credentials and API keys**
- ↗ **Rebuild contaminated RAG/DB** pipelines
- ↗ Verify **code integrity** via signing
- ↗ Document **response & regulatory compliance**

AI Incident Response Considerations:

- **Digital Investigations:** Investigators must have a clear understanding of AI systems and infrastructure to collect all relevant artifacts and know how to capture them.
- **Model Restoration:** If a model was compromised, redeploy a known-good version (from backups or version control). Retrain on sanitized data if needed.
- **Code and Pipeline:** If code was breached, revert to clean commits and rebuild from source. Purge any backdoors in scripts or configurations.
- **Credential Rotation:** Change all passwords, API keys, and tokens exposed in the incident. Check SaaS integrations and replace compromised credentials.
- **Data Integrity:** Validate and restore databases/knowledge stores. Rebuild any vector DBs with trusted documents to remove poisoned entries.
- **Infrastructure:** Redeploy container or VM images from fresh builds. Ensure no persistent artifacts remains (wipe, rebuild hosts if in doubt).
- **Hardening & Controls:** Post-incident, add stricter controls (e.g. model signing, stricter input validation). Review and tighten access policies.

Preparing for AI Incidents takes a multi faceted approach:

↗ Logging and Detection

- ↗ Log every model interaction: prompt, response, model version, embeddings, API calls.
- ↗ Integrate logs into SIEM/XDR platforms.
- ↗ Monitor for prompt anomalies, query spikes, or output drift.

↗ Authentication and Access

- ↗ Enforce short-lived API keys and OAuth tokens.
- ↗ Monitor token reuse or unexpected scopes.
- ↗ Use zero-trust network segmentation for model-serving endpoints.

↗ Containment and Recovery Playbooks

- ↗ Prompt Injection: Flush session context, revoke tokens, sanitize stored embeddings.
- ↗ Data Poisoning: Rebuild affected models from clean data; verify data lineage.
- ↗ Credential Compromise: Rotate tokens, disable compromised users, revalidate access control lists.
- ↗ Segment AI workloads (dedicated VMs/VPCs). If compromised, switch to backup model and environments, revoke tokens, and kill rogue sessions.

↗ Preparation and Testing

- ↗ Testing & Red Teams: Regularly test AI systems (adversarial testing, simulated poisonings) to validate defenses.
- ↗ AI Supply Chain Security: Sign and verify model artifacts; maintain Model Bills of Materials (MBOMs).
- ↗ Shadow AI Governance: Inventory all unapproved AI tools (browser extensions, SaaS bots).
- ↗ Cross-functional Coordination: Legal, engineering, and compliance teams must align on AI incident response.
- ↗ Training and Exercises: Conduct AI-specific tabletop simulations quarterly.
- ↗ Continuous Improvement: Use frameworks like MITRE ATLAS and OWASP Top 10 for LLMs to guide defense evolution.

Final Word on Incident Response for AI Systems

Key Takeaways	**Attack Surface**: AI systems contain new, multi-dimensional attack surface	**Forensics and Containment**: DFIR teams must evolve with AI telemetry & playbooks. AI incidents require retraining, not just patching	**Regulations**: Legal & regulatory frameworks are evolving and accelerating fast
Other Considerations	**Insurance & Planning**: Consider cyber insurance coverage for AI-specific breaches. Develop clear accountability (who is "owner" of AI assets).	**Ethical/Legal**: Stay alert for new laws on AI ethics and liability. Conduct Privacy/Impact Assessments for AI deployments (e.g. under GDPR).	**Fusion Alignment**: Governance and cross-functional alignment are critical
Evolving Risks	**Human Risk**: Educate users on safe AI usage. Enforce policies to prevent accidental data sharing with AI tools.	**Supply Chain**: Vet and secure third-party models (open-source or commercial) to prevent hidden vulnerabilities. Use model SBOMs and integrity checks.	**Future Trends**: Prepare for advances like autonomous AI agents; ensure IR plans evolve. Engage in industry collaboration on AI threat intel.

Thank you