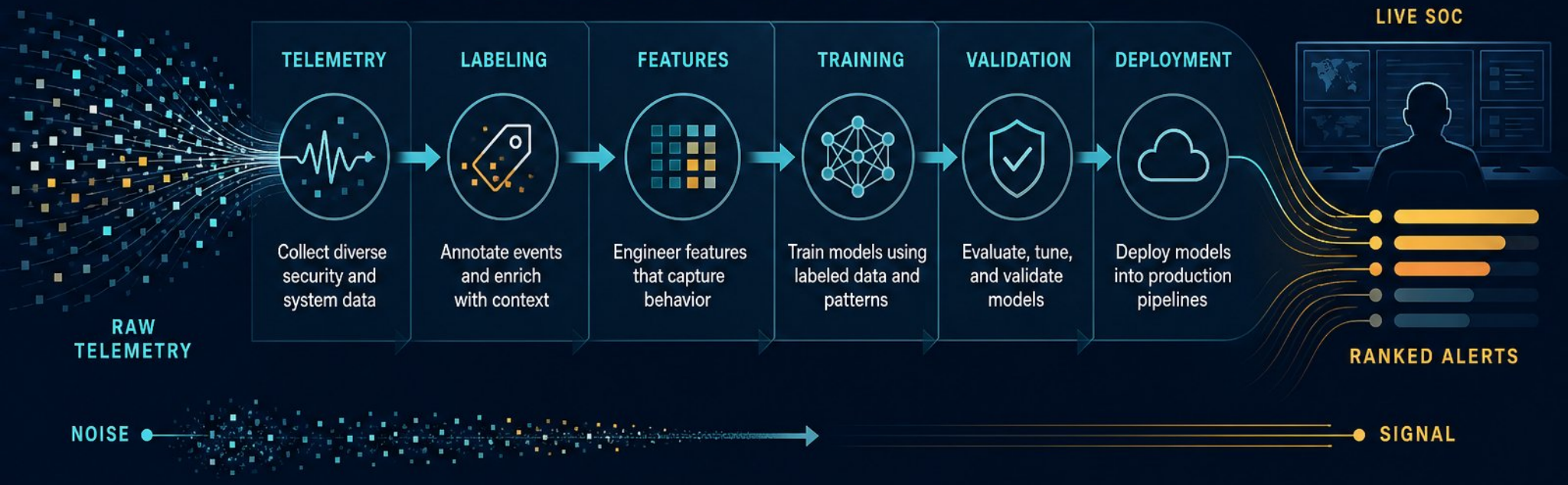
AI in Cybersecurity:
An Overview

- AI reshapes attack and defense: scale, speed, automation, attacker economics
- Dual-use dynamics: stronger detection and triage, but amplified phishing and evasion
- Roadmap: foundations, attacker AI, defensive AI, securing AI, governance, adoption
- Shared vocabulary: ML, LLMs, agents, SOC, SOAR, XDR, MTTD/MTTR
- 2026 context: adoption outran integration; governance and validation lag deployment

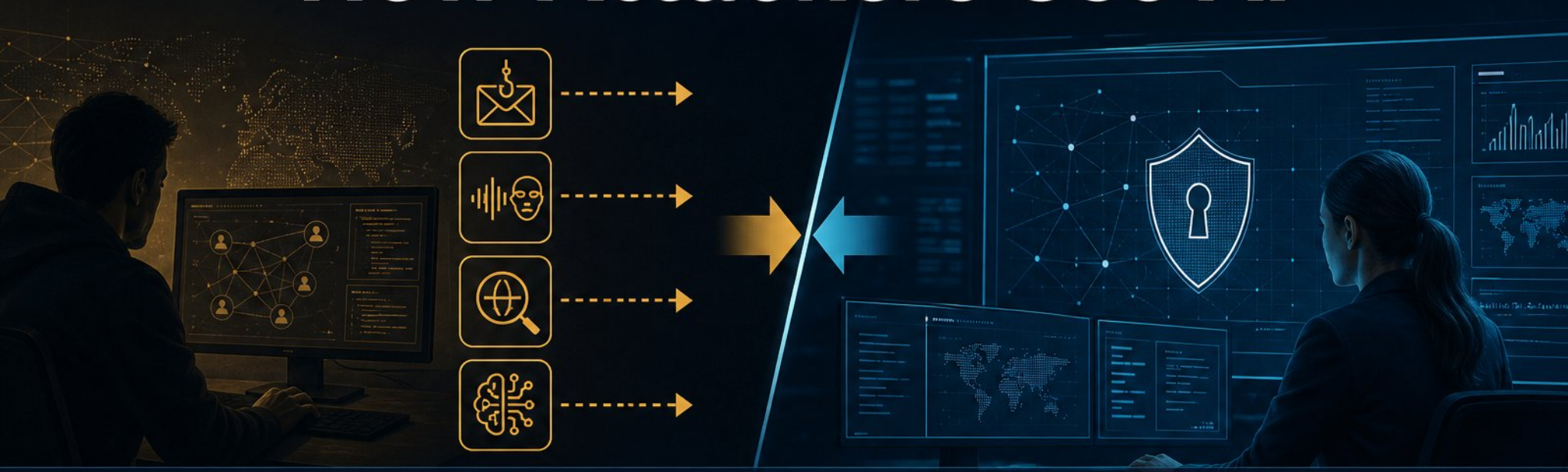
Background and Core Concepts

- Learning paradigms: supervised, unsupervised, semi-supervised, reinforcement
- Rule and signature detection versus ML anomaly detection; hybrid dominates
- Data pipeline: telemetry, labeling, features, validation, deployment
- Metrics: precision, recall, F1, false positives, base-rate effects
- Concept drift degrades models; retraining strategy matters as much as accuracy

SIGNAL-TO-NOISE: FROM TELEMETRY TO RANKED ALERTS



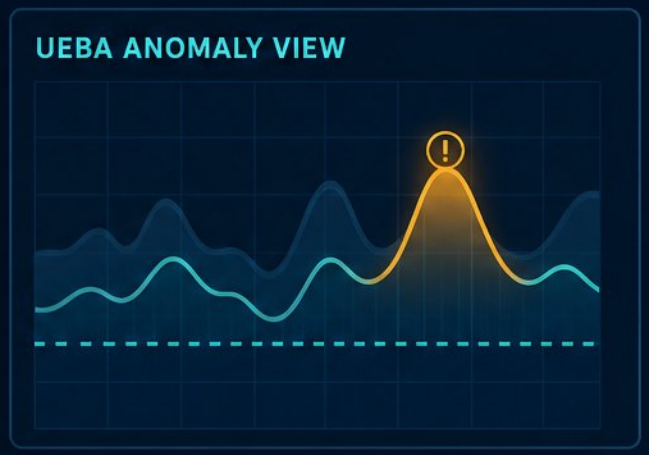
How Attackers Use AI

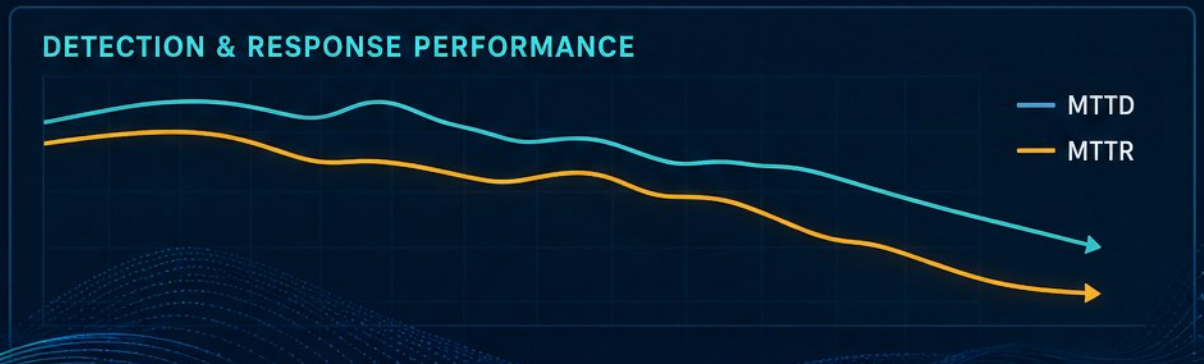


- AI spear phishing nearly triples click rates; personalization costs about $0.03 per email
- Deepfake voice and video fraud: 41% of organizations report audio-call social engineering
- Automated reconnaissance, vulnerability research, and adaptive malware speed up intrusion
- Prompt injection and adversarial evasion target ML defenses and AI-enabled tooling
- Constraints remain: cost, reliability, detection risk, and human direction still needed

AI for Detection and Response

- Behavioral analytics, UEBA, and anomaly detection catch what rules miss
- AI triage enriches, correlates, and prioritizes alerts before analysts see them
- SOAR and XDR automate containment: isolation, blocking, credential revocation
- Measured gains: 75–99% fewer manual reviews, MTTR cut by 50%+
- AI triage is not autonomous judgment: evidence trails and human review stay essential





LLMs, Agents, and the Agentic SOC

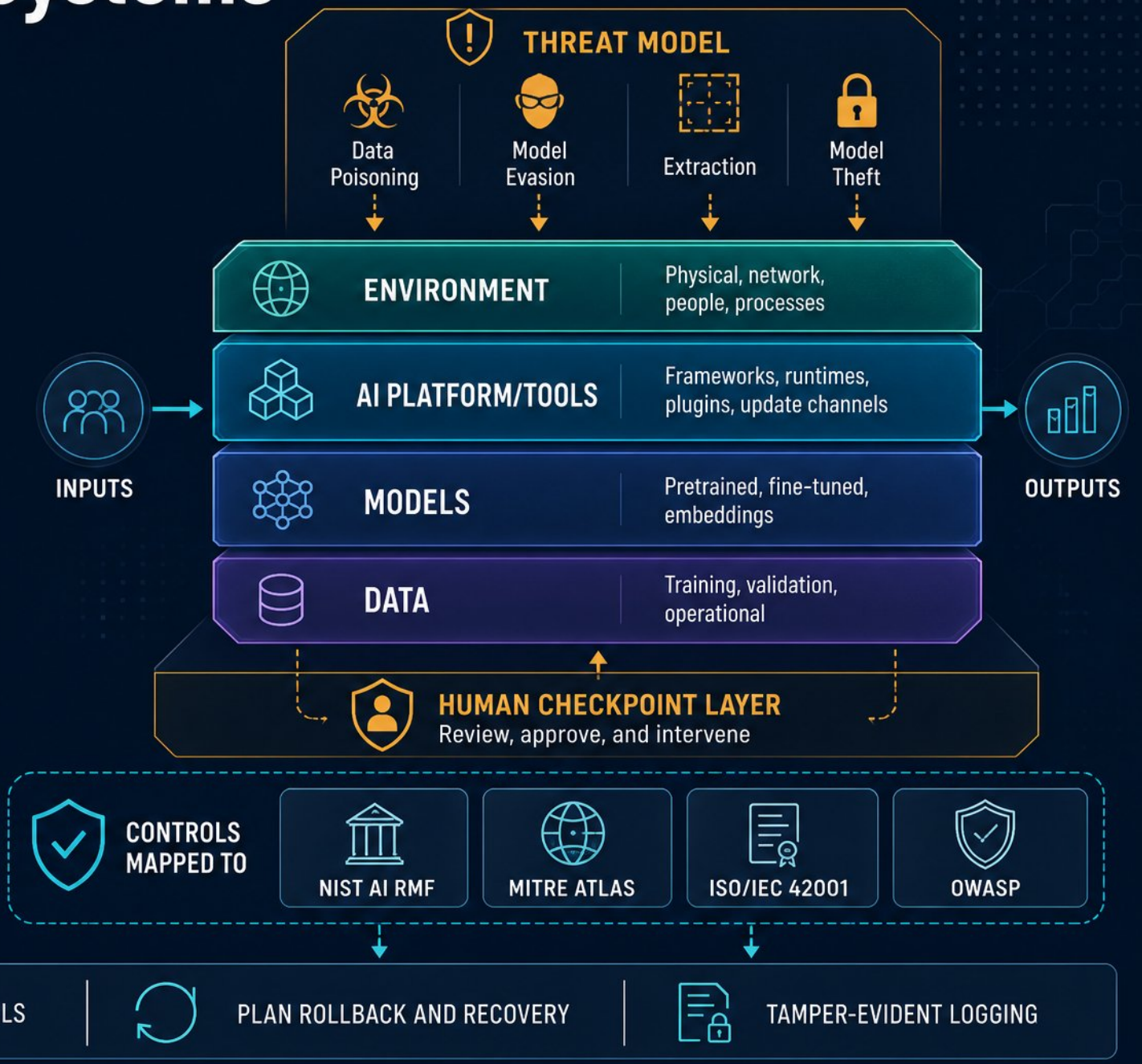
- **LLM use cases:** summarization, log analysis, investigation drafting
- **Copilots and bounded agents:** scoped tasks, tools, memory, orchestration
- **Why adopt:** alert volume, speed, consistency, 24/7 coverage gaps
- **Architecture:** phase-scoped agents, tool orchestration, human-in-the-loop gates
- **Reality:** agents do repeatable work; humans own judgment and escalation

Agentic AI Risks and Guardrails

- Security agents expand the attack surface: goal hijack, tool misuse, identity abuse, memory poisoning
- Indirect prompt injection flows from untrusted logs, email, and threat intelligence feeds
- Tool orchestration and memory are primary trust boundaries; scope them tightly
- Guardrails: least privilege, scoped identities, sandboxing, input validation, output monitoring
- Deployment gate: human override tabletop, circuit breakers, and go/no-go criteria before scaling

Securing AI Systems

- - Threat modeling covers data poisoning, model evasion, extraction, and model theft
- - Third-party models, plugins, and update channels create supply chain and provenance gaps
- - Map AI controls to NIST AI RMF, MITRE ATLAS, ISO/IEC 42001, and OWASP
- - Protect four elements: environment, AI platform/tools, models, and data
- - Monitor misbehaving models; plan rollback, recovery, and tamper-evident logging



Governance, Risk, and Compliance

EU AI Act enforceable
since August 2, 2026: transparency and GPAI active

High-risk obligations
delayed: Annex III to December 2027

High-risk obligations
delayed: Annex I to August 2028

Map to NIST AI RMF

Map to ISO/IEC 42001

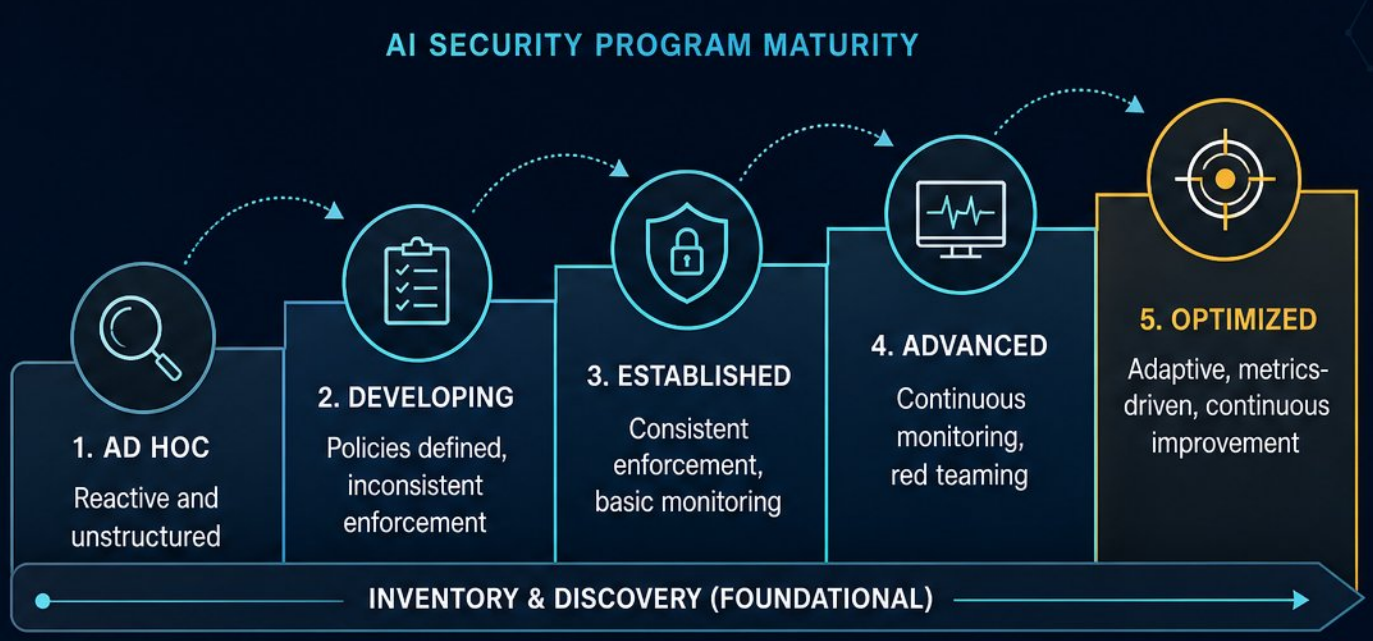
Map to MITRE ATLAS

Map to OWASP LLM Top 10

- Risk assessment, documentation, and AI risk registers for security tools
- Evidence: logs, test records, explainability, vendor due diligence

AI Security Program Maturity and Prioritization

- Inventory first: you cannot secure AI systems you have not enumerated
- Five levels: ad hoc to optimized—policy, enforcement, monitoring, red teaming
- Prioritize by risk and feasibility: data exposure, business impact, attacker interest
- Most SOCs: start with vendor capabilities, customize where justified
- Track coverage, validation, false-positive load, time-to-respond, control effectiveness



Practical Playbook and Adoption Roadmap

- Start with bounded, high-volume, low-variance cases: phishing triage, enrichment, summarization
- Pilot design: clear baselines, success metrics, elevated logging, defined exit criteria
- Redesign workflows, escalation thresholds, and analyst responsibilities around AI-assisted investigation
- Common failure modes: unstructured use, no validation, integration gaps, automation without governance
- Scale only with demonstrated accuracy, human trust, auditability, and measurable risk reduction

Dead Ends: Risk of Poor Outcomes and Rework

- **Context:** 79% use AI, only 36% have integrated it into a defined SOC workflow

Hands-On Considerations and Safe Experimentation

- Lab ideas: prompt injection, LLM red teaming, secure RAG pipelines, agent tool misuse
- Run open-source models locally to keep sensitive security data contained
- Validate with labeled benchmarks, red-team exercises, and continuous comparison
- Tabletop AI incidents: deepfake fraud, agent misuse, model compromise
- Avoid unsafe tests: sandboxing, data handling rules, approval gates for offensive work

Workforce, Roles, and Team Readiness

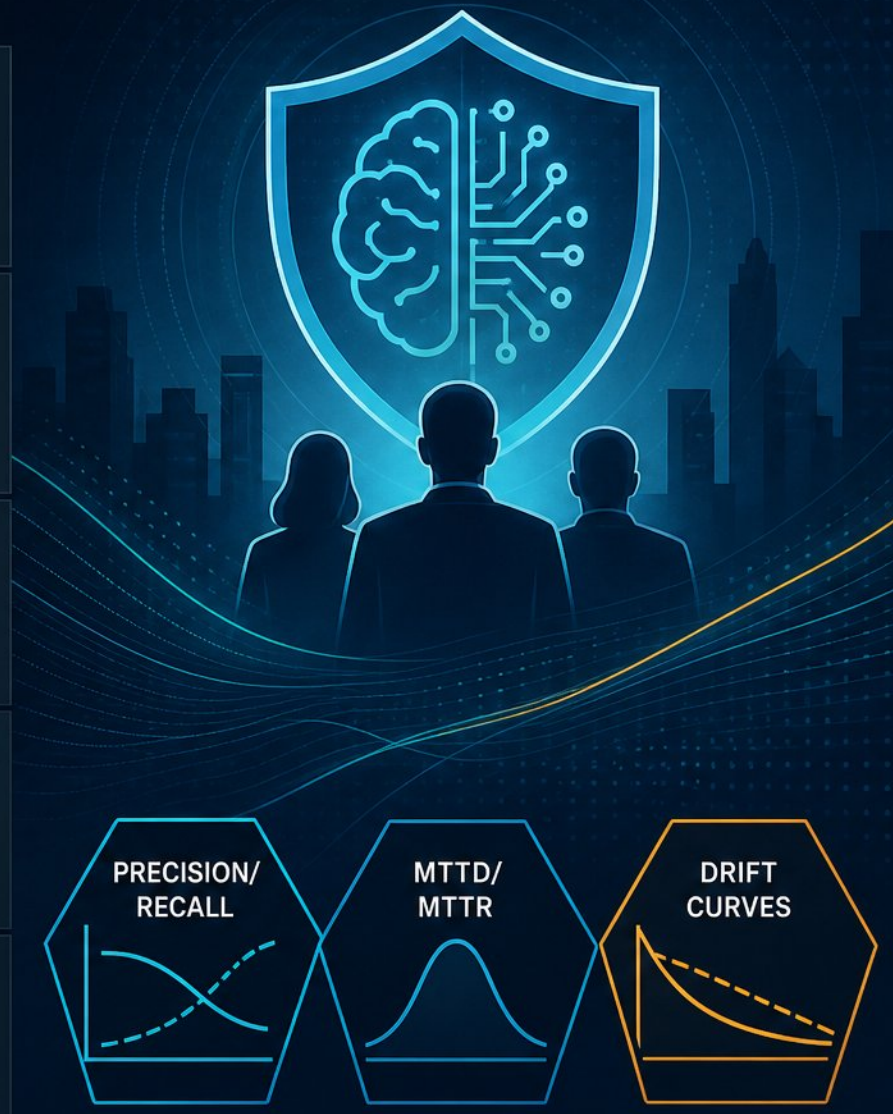
- Analysts supervise AI outcomes; detection engineers teach the system
- SOC leaders orchestrate autonomy, escalation, and accountability
- Skills gaps: data literacy for analysts, AI security for developers
- Vendor-neutral certs, hands-on labs, role-aligned curricula
- AI security extends existing roles, not a separate job family
- Adoption has outpaced integration; training demands rose sharply in 2026

Looking Ahead:
Trends and Strategic Decisions

- AI-enabled attacks become the norm, not the exception, through 2027
- Emerging capabilities: autonomous defense, agentic identity management, machine-speed response
- Expanded attack surface: AI agents, supply chains, deepfakes, critical infrastructure, physical AI
- Market signals: securing-AI spending jumps to $4.8B in 2027
- Next 12–24 months: what to invest in, pilot, and monitor

Course Wrap-Up and Decision Checklist

- **Recap:** dual-use AI, agentic SOC, securing AI systems, governance, workforce readiness
- **Immediate actions:** inventory AI usage, validate outputs, establish guardrails, update playbooks
- **30/90/365-day roadmap:** baseline, pilot, harden, scale, reassess
- **Open questions:** agent identity, model provenance, explainability, regulatory harmonization
- **Leader checklist:** risk appetite, vendor due diligence, human oversight, measurable outcomes



PersonWise.

PersonWise • AI digital-human course platform

Scan the code or click anywhere to watch this course live, with voice Q&A

personwise.ai/t/268cde5a/public