David Pierce, MPA

BLUF AI/ML Observability & Safety Teams are key to organizational health.

Executive Summary: Enterprise AI/ML Observability & Safety teams are crucial in mitigating emerging AI-specific risk in a compliant way. Solutions available in the market are immature and fail to effectively manage Cyber[1] & Legal[2] exposure. State-backed[3] actors are actively targeting market leaders[4], while unpatchable[5] Gen AI vulnerabilities add to existing adversarial attacks against ML systems. Detailed below are AI-specific risks, relevant mitigations, as well as the skills necessary for these teams; core products and services are similarly annotated.

Emerging Exposure:

- Regulatory – EU AI Act[6], WH Executive Order[7], and Pending Legislation
- Reputational – Negative Customer Experiences, Hallucinations[8], etc
- Data Loss – Extracting Training[9], Contextual[10], or Decoding Encrypted Customer Data[11]
- External IP Theft – Decompilation[12] of Distributed Applications (e.g. App Stores)
- Hidden Intrusions – Difficult to Detect AI Backdoors[13]; Persistant Despite Safety Training

Per-Risk Mitigations:

- Regulatory – ‘High-Risk’ Controls, Documented Process, Internal Audit
- Reputational – Real-Time Monitoring, Contextual Grounding, etc
- Data Loss – Event-Driven DLP, Customer-Managed Encryption Keys, etc
- IP Theft – AI-aware Application Architecture & Code Obfuscation
- Hidden Intrusions – Sanitized Inputs & Versioned Sources of Truth

Core Competencies:

- External Observability - ‘Managing Alerts at-Scale by Exception’
- Internal Explainability - ‘Detailed Audit of AI Decisions’
- Per-Component Securability - ‘Minimizing Residual Risk’

Core Products:

- Per-Application Baselines w/ Configurable Thresholds & Pre-Deployment Automation
- Real-Time Alerting via Event-Driven Interdiction Suite for Deployed Assets
- Automated Playbooks for AI Vulnerability Testing (e.g. Kill Chain & Defense Plan)
- Customizable Instrumentation via Transformer Heads using Mechanistic Interpretability

Core Services:

- AI Software Integration & Strategic Planning
- Data Pipeline & Platform Modernization
- Cross-Domain AI Upskilling for Existing Teams
- Audit & Regulatory Support for Business

Citations:

1) NIST Vulnerability Report
2) Improperly Licensed AI Training Data
3) State-Backed AI-Security Efforts
4) Market Leaders Compromised
5) LLM Refusal Mediated by Single Direction
6) EU AI Act Explained
7) WH Executive Order
8) AI Hallucination Explained
9) Extracting Training Data
10) Inverting AI Embeddings
11) Decoding Encrypted AI Messages
12) Decompilation & Extraction of IP
13) Sleeper Agents That Persist