**Product Requirements Document: AI Security Posture Management (AI-SPM) Platform**

**1. Introduction**

This document outlines the product requirements for an AI Security Posture Management (AI-SPM) Platform. The platform is designed to provide organizations with a comprehensive, continuous, and proactive approach to securing their Artificial Intelligence (AI) and Machine Learning (ML) systems, the data they consume and generate, and the underlying infrastructure. As AI/ML adoption accelerates, traditional security measures are insufficient to address the unique and evolving vulnerabilities inherent in these complex systems, such as data poisoning, model evasion, prompt injection, and risks associated with their often opaque "black-box" nature. This AI-SPM platform aims to empower organizations to innovate responsibly and leverage AI's full potential by embedding security throughout the AI lifecycle.

**2. Goals and Objectives**

* **Enable Secure AI Innovation:** Allow organizations to confidently develop, deploy, and scale AI initiatives by providing robust security foundations.
* **Proactive Risk Mitigation:** Systematically identify, assess, prioritize, and mitigate AI-specific security risks and vulnerabilities before they can be exploited.
* **Ensure Comprehensive Compliance:** Facilitate adherence to evolving global AI regulations, industry standards, and data privacy mandates (e.g., NIST AI RMF, ISO/IEC 23894, ENISA FAICP, GDPR, CCPA).
* **Build Trust and Transparency:** Enhance stakeholder confidence in AI systems by ensuring their security, integrity, fairness, and reliability.
* **Improve Operational Efficiency:** Streamline AI security operations through automation, centralized visibility, and actionable insights.

**3. Target Audience**

* **Business Stakeholders:**
  + Chief Information Security Officers (CISOs)
  + Chief Risk Officers (CROs)
  + Data Protection Officers (DPOs)
  + Compliance and Governance Officers
  + Business Leaders and Executives driving AI initiatives
* **Technical Stakeholders:**
  + Security Operations (SecOps) Teams
  + AI/ML Engineers and Developers
  + Data Scientists
  + MLOps Engineers
  + DevSecOps Teams
  + Cloud Security Architects
  + IT Administrators

**4. User Stories**

* **US1 (CISO):** As a CISO, I want a unified dashboard providing a real-time overview of our organization's AI security posture and compliance status, so I can effectively manage risk and report to the board.
* **US2 (Security Analyst):** As a Security Analyst, I want to continuously monitor AI models for anomalous behavior and known attack patterns (like prompt injection or data poisoning), so I can rapidly detect and respond to threats.
* **US3 (AI/ML Engineer):** As an AI/ML Engineer, I want to integrate automated security checks and vulnerability scanning into our MLOps pipelines, so I can build and deploy secure AI models from the outset.
* **US4 (Data Scientist):** As a Data Scientist, I want tools to verify the integrity of training datasets and protect against data manipulation, so our models produce accurate and reliable results.
* **US5 (Compliance Officer):** As a Compliance Officer, I want to generate comprehensive audit logs and compliance reports for our AI systems against frameworks like NIST AI RMF and GDPR, so I can demonstrate adherence to regulatory requirements.
* **US6 (MLOps Engineer):** As an MLOps Engineer, I want to securely configure and deploy AI models into production environments, ensuring their integrity and preventing unauthorized access or modifications.
* **US7 (Cloud Security Architect):** As a Cloud Security Architect, I want to discover all AI assets across our multi-cloud environments and identify potential attack paths to these assets, so I can design and implement effective security controls.

**5. Product Features and Requirements**

**5.1. Functional Requirements**

**FR1: Visibility, Discovery, and AI Inventory Management** \* FR1.1: **Automated AI Asset Discovery:** Continuously discover and inventory all AI assets, including models (custom, open-source, third-party), data sources (training, inference, RAG), data pipelines, APIs, underlying infrastructure components (cloud, on-prem), and dependencies across the organization. \* FR1.2: **AI Bill of Materials (AIBOM):** Generate and maintain an AIBOM for each AI system, detailing its components, libraries, data lineage, and dependencies, aligning with standards like OWASP CycloneDX. \* FR1.3: **Asset Classification:** Automatically classify AI assets based on factors such as business criticality, data sensitivity (e.g., presence of PII/PHI), and potential impact if compromised. \* FR1.4: **Shadow AI Detection:** Identify and flag unauthorized or unmanaged AI models and applications ("shadow AI") operating outside governance frameworks. \* FR1.5: **Ecosystem Visualization:** Provide an interactive map or graph (e.g., "Security Graph") visualizing the AI ecosystem, its components, data flows, and interdependencies to understand potential attack paths.

**FR2: Data Governance and Protection in AI Systems** \* FR2.1: **Sensitive Data Discovery & Classification:** Automatically scan and identify sensitive data (e.g., PII, PHI, financial data, intellectual property) within AI training datasets, model inputs/outputs, vector databases, and operational logs. \* FR2.2: **Data Privacy Policy Enforcement:** Enable configuration and enforcement of data privacy policies (e.g., GDPR, CCPA, HIPAA) specific to AI systems, including consent management and data subject rights. \* FR2.3: **Privacy-Enhancing Technology (PET) Support:** Support or integrate with techniques like data anonymization, pseudonymization, differential privacy, and federated learning to protect data used in AI. \* FR2.4: **Data Flow Monitoring & Control:** Monitor and control data flows to, from, and within AI systems to prevent unauthorized data movement or exposure. \* FR2.5: **Data Leakage Prevention (DLP) for AI:** Detect and prevent leakage of sensitive information through model outputs, inference attacks (membership/attribute inference), or unintentional disclosures by generative AI. \* FR2.6: **Secure Data Lifecycle Management:** Enforce secure data handling practices throughout the AI data lifecycle: secure collection, pre-processing, labeling, storage (encryption at rest/in-transit), access controls (least privilege), and secure deletion.

**FR3: Comprehensive AI Risk Management** \* FR3.1: **AI Model Vulnerability Scanning:** Scan AI models (supporting formats like H5, Pickle, ONNX, SavedModel) for known vulnerabilities, unsafe code, malicious implants, backdoors, and serialization issues. \* FR3.2: **AI Threat Modeling:** Provide tools or integrations for AI-specific threat modeling, aligning with frameworks like MITRE ATLAS to identify potential attack vectors. \* FR3.3: **Contextual Risk Assessment & Prioritization:** Assess and prioritize AI-related risks based on vulnerability exploitability, asset criticality, business impact, data sensitivity, and real-time AI-specific threat intelligence. \* FR3.4: **Misconfiguration Detection & Management:** Identify and report misconfigurations in AI platforms, cloud services supporting AI, data storage, and network settings. \* FR3.5: **AI Supply Chain Security:** Assess risks associated with third-party AI models, pre-trained foundation models, open-source libraries, APIs, and data sources used in the AI supply chain. \* FR3.6: **AI Attack Path Analysis:** Identify and visualize potential attack paths to critical AI assets and sensitive data.

**FR4: Runtime Monitoring, Threat Detection, and Response** \* FR4.1: **Real-Time AI Behavior Monitoring:** Continuously monitor AI model behavior, API interactions, user prompts (especially for LLMs), and model outputs in real-time. \* FR4.2: **AI Anomaly Detection:** Establish baselines for normal AI system behavior and detect deviations or anomalies that could indicate a security threat or operational issue. \* FR4.3: **Adversarial Attack Detection & Prevention:** Detect and mitigate adversarial attacks, including data poisoning, model evasion, and membership/attribute inference attacks. \* FR4.4: **Model Theft/Extraction Detection:** Identify patterns indicative of model extraction or intellectual property theft attempts. \* FR4.5: **Model Integrity Verification:** Continuously verify the integrity of deployed AI models to ensure they have not been tampered with or unauthorizedly modified. \* FR4.6: **Automated Incident Response:** Provide customizable, automated response workflows for AI-related security incidents (e.g., alerting, quarantining models, blocking malicious IPs, revoking access, initiating patching). \* FR4.7: **AI Firewall Capabilities:** Implement AI-specific firewall functionalities to inspect and filter inbound queries and outbound responses for AI applications and LLMs, blocking malicious inputs (e.g., prompt injection, SQL injection via prompts), unauthorized queries, data scraping, and harmful outputs.

**FR5: Securing the AI Lifecycle (DevSecOps/MLSecOps Integration)** \* FR5.1: **MLOps & DevSecOps Integration:** Seamlessly integrate with popular MLOps platforms (e.g., Azure Machine Learning, Kubeflow, MLflow, Vertex AI, SageMaker) and CI/CD pipelines to embed security throughout the AI development lifecycle. \* FR5.2: **Lifecycle Security Controls:** Provide security controls and checks at each stage of the AI lifecycle: data acquisition, preprocessing, model development, training, validation, testing, deployment, operation, and decommissioning. \* FR5.3: **Secure Development Practices:** Enforce secure coding practices for AI-related software, scan for vulnerabilities in AI code and dependencies (SAST/SCA for AI). \* FR5.4: **Secure Configuration Management:** Manage and audit configurations for AI models, platforms, and infrastructure to prevent drift and ensure secure defaults. \* FR5.5: **Model Versioning and Integrity Tracking:** Track model versions and ensure integrity across different stages of the lifecycle.

**FR6: AI Agent Security (for systems with autonomous AI agents)** \* FR6.1: **Prompt Filtering & Validation:** Implement mechanisms to filter, validate, and sanitize prompts sent to AI agents, blocking malicious or out-of-policy inputs and preventing prompt injection attacks. \* FR6.2: **RAG Data Protection & Access Control:** Secure data sources used by Retrieval-Augmented Generation (RAG) pipelines, enforce access controls on retrieved data, and filter sensitive information before it reaches the model or output. \* FR6.3: **Secure External Action Control:** Govern actions AI agents can perform on external systems (e.g., API calls, database modifications, transactions). Enforce authorization policies, track "on behalf of" chains for accountability, and allow for human-in-the-loop approval for high-risk actions. \* FR6.4: **AI Agent Response Enforcement:** Monitor and filter responses generated by AI agents for accuracy, compliance, sensitivity, and appropriateness before delivery to users or other systems. Prevent leakage of PII, credentials, or internal knowledge.

**FR7: Governance, Compliance, and Reporting** \* FR7.1: **AI Governance Policy Management:** Provide a centralized platform to define, manage, and enforce AI security, ethics, and governance policies across the organization. \* FR7.2: **Automated Compliance Mapping & Monitoring:** Continuously monitor and assess AI systems against relevant regulatory frameworks and standards (e.g., NIST AI RMF, ISO/IEC 23894, ENISA FAICP, GDPR, CCPA, HIPAA, PCI-DSS). \* FR7.3: **Audit Trails & Reporting:** Generate comprehensive audit trails for all AI security-related activities, model lineage, approvals, and risk acceptance. Provide customizable dashboards and exportable reports for compliance and stakeholder communication. \* FR7.4: **Role-Based Access Control (RBAC):** Implement granular RBAC for the AI-SPM platform itself, ensuring users only have access to features and data relevant to their roles. \* FR7.5: **AI Red Teaming Support:** Facilitate AI red teaming exercises by providing insights, logging capabilities, and integration points for red teaming tools (e.g., Garak, PyRIT, AI Fairness 360, Foolbox).

**FR8: Platform Capabilities** \* FR8.1: **Scalable Architecture:** Design the platform with a scalable architecture to accommodate a growing number of AI models, data volumes, and user activity. \* FR8.2: **Broad Integrations:** Ensure seamless integration with major cloud platforms (AWS, Azure, GCP), on-premises environments, common AI/ML development tools, MLOps platforms, and existing security ecosystem tools (SIEM, SOAR, EDR, CNAPP). \* FR8.3: **User Experience (UX):** Provide an intuitive, user-friendly interface with customizable dashboards, clear visualizations, and actionable alerts for diverse user roles. \* FR8.4: **Confidential Computing Support:** Integrate with or support confidential computing technologies (e.g., secure enclaves, TEEs) to enhance the privacy and integrity of AI model training and inference, especially for sensitive data. \* FR8.5: **AI-Specific Threat Intelligence:** Incorporate or integrate with up-to-date AI-specific threat intelligence feeds to inform risk assessments and detection capabilities.

**5.2. Non-Functional Requirements**

* **NFR1: Performance:** The platform must perform real-time monitoring, detection, and response with minimal latency and negligible impact on the performance of monitored AI systems.
* **NFR2: Scalability:** The system must be able to scale horizontally and vertically to support a large and growing number of AI assets, users, and data throughput.
* **NFR3: Reliability & Availability:** The platform must be highly reliable and available (e.g., 99.99% uptime), with robust fault tolerance and disaster recovery mechanisms.
* **NFR4: Usability:** The interface should be intuitive and accessible for users with varying levels of technical expertise, providing clear, actionable information.
* **NFR5: Security:** The AI-SPM platform itself must adhere to the highest security standards, protecting its own data and configurations from unauthorized access or tampering.
* **NFR6: Interoperability:** The platform must easily integrate with a diverse ecosystem of third-party tools and platforms via APIs and standard protocols.
* **NFR7: Maintainability:** The platform should be designed for ease of maintenance, updates, and upgrades.
* **NFR8: Extensibility:** The platform should allow for the addition of new detection rules, compliance frameworks, and integrations as the AI security landscape evolves.

**6. Success Metrics**

* **Security Effectiveness:**
  + Reduction in the number and severity of AI-related security incidents.
  + Mean Time to Detect (MTTD) and Mean Time to Respond (MTTR) for AI-specific threats.
  + Percentage of AI assets covered and actively monitored by the AI-SPM platform.
  + Reduction in identified AI vulnerabilities and misconfigurations over time.
* **Compliance & Governance:**
  + Improvement in compliance scores against relevant AI regulations and standards.
  + Reduction in audit findings related to AI security.
  + Percentage of AI systems with complete AIBOMs and data lineage documentation.
* **Operational Efficiency:**
  + Time saved in AI security assessments and compliance reporting through automation.
  + Reduction in manual effort for AI threat hunting and incident response.
* **User Adoption & Satisfaction:**
  + Number of active users and teams utilizing the platform.
  + User satisfaction scores (e.g., Net Promoter Score - NPS, user surveys).
  + Reduction in instances of "shadow AI" due to improved visibility and governance.
* **Risk Reduction:**
  + Overall improvement in the organization's AI security posture score (if quantifiable).
  + Reduction in potential financial impact from AI-related risks.

**7. Assumptions and Constraints**

**7.1. Assumptions:** \* Organizations recognize the unique security risks associated with AI/ML and are willing to invest in specialized solutions. \* Target organizations have a growing portfolio of AI/ML models and applications in development or production. \* Access to AI system logs, configurations, and model details will be available for monitoring and assessment (with appropriate permissions). \* Collaboration between security, AI/ML, data, and business teams is achievable for effective AI-SPM implementation.

**7.2. Constraints:** \* The AI threat landscape and AI technologies are rapidly evolving, requiring continuous updates and adaptation of the platform. \* Integrating with a highly diverse and fragmented ecosystem of AI development tools, MLOps platforms, and cloud services presents complexity. \* Potential scarcity of specialized AI security expertise within customer organizations may impact adoption and effective use. \* The global regulatory landscape for AI is still developing and varies by jurisdiction, requiring flexibility in compliance modules. \* Performance overhead of monitoring and security controls on AI systems must be minimized.

**8. Future Considerations (Potential Roadmap Items)**

* **Advanced Predictive AI for Threat Anticipation:** Leverage AI within the AI-SPM platform to predict emerging AI attack vectors and proactively recommend countermeasures.
* **Explainable AI (XAI) Integration for Security:** Integrate XAI techniques to provide deeper insights into why an AI model is flagged as vulnerable or behaving anomalously, aiding in remediation.
* **Automated Secure AI Development Guidance:** Provide developers with real-time, context-aware security recommendations and best practices directly within their AI development environments.
* **Enhanced Support for Federated Learning & Homomorphic Encryption:** Deeper integrations and management capabilities for AI systems utilizing advanced privacy-enhancing technologies.
* **Decentralized/Blockchain-based AI Trust & Auditability:** Explore features for enhanced trust and immutable audit trails for AI operations using distributed ledger technology.
* **AI-Driven Security Policy Generation:** Assist organizations in automatically generating tailored AI security policies based on their specific AI use cases, risk appetite, and regulatory obligations.
* **Community-Driven Threat Intelligence & Defense Sharing:** Facilitate an anonymized and secure platform for sharing AI threat intelligence and effective defense strategies among users.

This PRD provides a foundational blueprint for developing a robust AI Security Posture Management platform. It should be considered a living document, subject to refinement as market needs and technological capabilities evolve.