**Model Training for Data Security and Incident Response Policy (MTDS-IRP v1.0)**

**Classification:** Internal Use / Confidential  
**Policy Owner:** Chief Information Security Officer (CISO) Keitavius Alexander  
**Department:** AI Governance, Security & Compliance  
**Effective Date:** 10/16/2025  
**Last Reviewed:** 10/16/2025  
**Next Review Date:** 10/16/2026  
**Approved By:** AI Governance and Risk Oversight Committee

**Table of Contents**

1. **Executive Summary**
2. **Purpose**
3. **Scope**
4. **Policy Statement**
5. **Roles and Responsibilities**
6. **Policy Requirements**
   * 6.1 AI-Specific Threat Protections
   * 6.2 Insider Threat Controls
   * 6.3 Data Lifecycle and Provenance
   * 6.4 Device and Endpoint Security
   * 6.5 Network Security
   * 6.6 Access Control and Least Privilege
   * 6.7 Security Awareness and Training
   * 6.8 Data Privacy and Regulatory Compliance
   * 6.9 Model Integrity and Verification
   * 6.10 Vendor and Third-Party Controls
7. **Incident Response Procedures**
   * 7.1 Incident Reporting
   * 7.2 Incident Classification and Severity
   * 7.3 Response Procedures
   * 7.4 Communication and Disclosure
8. **Enforcement and Sanctions**
9. **References and Framework Alignment**
10. **Appendices**
    * 10.1 Key Definitions and Acronyms
    * 10.2 Version Control and Approval Record
    * 10.3 Regulatory and Framework Alignment Map
    * 10.4 Sample Incident Reporting Workflow
    * 10.5 Policy Intent and Summary

**1. Executive Summary**

The Model Training for Data Security and Incident Response Policy (MTDS-IRP) establishes a comprehensive governance framework for protecting the confidentiality, integrity, and availability of data and systems used in AI model development.

It addresses the full spectrum of AI-specific threats, including data poisoning, model inversion, insider risk, and adversarial attacks, while enforcing robust incident response protocols.  
This policy ensures alignment with NIST AI RMF, ISO/IEC 23894, ISO/IEC 27001, and SOC 2 standards, and supports the organization’s broader GRC and Zero Trust initiatives.

**2. Purpose**

The purpose of this policy is to:

* Define mandatory security and governance controls for AI model training environments.
* Protect datasets, models, and training pipelines from intentional or accidental compromise.
* Enforce consistent incident response procedures across all AI systems.
* Maintain regulatory compliance and uphold ethical AI principles.
* Support audit readiness through verifiable logging, documentation, and reporting.

**3. Scope**

This policy applies to:

* All personnel (employees, contractors, consultants) with access to AI training data or systems.
* All technology assets, including servers, endpoints, networks, and cloud services supporting AI training and validation.
* Third-party vendors and partners providing AI-related datasets, APIs, or processing tools.

It governs data handling, model lifecycle management, access control, and incident response for all AI systems operated by the organization.

**4. Policy Statement**

The organization is committed to securing AI training environments through a unified governance framework that integrates information security, privacy, and ethical AI risk management.

All personnel must comply with this policy and its associated standards. Violations may result in disciplinary action or termination of access and contracts.

**5. Roles and Responsibilities**

* **Executive Leadership:** Approves this policy, allocates resources, and ensures enterprise-wide enforcement.
* **Security & Compliance Team:** Implements security controls, monitors adherence, conducts audits, and manages incident response.
* **IT Operations:** Maintains endpoint, network, and cloud infrastructure security through encryption, patching, and monitoring.
* **AI Trainers / Annotators:** Follow security protocols, complete assigned training, and report any anomalies or security incidents.
* **Third-Party Vendors:** Comply with Data Security Agreements (DSAs) and undergo continuous security vetting.

**6. Policy Requirements**

**6.1 AI-Specific Threat Protections**

* Protect against data poisoning, model inversion, membership inference, prompt injection, and adversarial attacks.
* Implement differential privacy, output filtering, and restricted model access to prevent data leakage.
* Conduct periodic red-team exercises, integrity validation, and stress testing on datasets and models.

**6.2 Insider Threat Controls**

* Mandatory background checks for all personnel engaged in AI model training.
* All work must occur on company-managed devices or secure VDIs.
* Enable keystroke and activity monitoring, session logging, and behavioral anomaly detection.
* Escalate suspicious activities immediately (e.g., off-hour access, mass data downloads, policy bypass attempts).

**6.3 Data Lifecycle and Provenance**

* Maintain complete metadata for every dataset: source, ownership, version, and modification history.
* Sandbox unverified data before use; only verified datasets enter model training pipelines.
* Maintain audit trails for preprocessing, cleaning, and model training actions.
* Conduct periodic dataset bias reviews and remove unapproved or malicious content.

**6.4 Device and Endpoint Security**

* Only organization-issued devices or secure VDIs may access AI systems.
* Enforce AES-256 full-disk encryption, secure boot, and endpoint detection and response (EDR).
* Ensure automated patch management and continuous compliance monitoring.
* Require Multi-Factor Authentication (MFA) for all user accounts.

**6.5 Network Security**

* Require VPN usage for all external connections; split tunneling is prohibited.
* Public Wi-Fi use is forbidden without an approved secure VPN connection.
* Enforce firewalls, IDS/IPS, and Zero Trust segmentation across all environments.
* Implement continuous network traffic monitoring and alerting.

**6.6 Access Control and Least Privilege**

* Apply Role-Based Access Control (RBAC) for datasets, model repositories, and tools.
* Prohibit shared accounts; rotate credentials regularly.
* Maintain comprehensive access logs for at least 12 months and review them quarterly.

**6.7 Security Awareness and Training**

* All personnel must complete mandatory onboarding and quarterly refresher training.
* Training must cover AI-specific threats, secure data handling, incident reporting, and phishing prevention.
* All personnel must sign an Acceptable Use and Confidentiality Agreement before accessing production data or models.

**6.8 Data Privacy and Regulatory Compliance**

* Personal or sensitive data must be anonymized or pseudonymized.
* Compliance with GDPR, HIPAA, CCPA, and applicable local laws is mandatory.
* Data retention, archival, and deletion must follow approved schedules aligned with legal obligations.

**6.9 Model Integrity and Verification**

* Validate all models using gold-standard datasets prior to deployment.
* Monitor models for bias, anomalies, or unauthorized changes.
* Conduct periodic adversarial testing to confirm robustness and resilience.

**6.10 Vendor and Third-Party Controls**

* Vendors must undergo security vetting, DPA/DSA signing, and continuous compliance checks.
* Vendors must meet internal device, network, and access security requirements.
* Perform annual third-party audits to confirm adherence to this policy.

**7. Incident Response Procedures**

**7.1 Incident Reporting**

* All employees must immediately report suspected incidents, including unauthorized access, data poisoning, model corruption, or malware infection.

**7.2 Incident Classification and Severity**

**Incidents must be classified by severity level to determine escalation, response time, and containment procedures. The following categories apply:**

* **Low Severity:**
  + Examples: Contained malware, minor policy violations, low-impact technical anomalies.
  + Response Priority: Must be investigated and remediated within 24 hours.
* **Medium Severity:**
  + Examples: Unauthorized access attempts, dataset corruption, temporary service disruption.
  + Response Priority: Investigation must begin within 12 hours, with containment and recovery initiated immediately after verification.
* **High Severity:**
  + Examples: Confirmed breach, data exfiltration, model compromise, insider threat, or regulatory-impacting event.
  + Response Priority: Immediate escalation to the Incident Response Team (IRT), with full containment and executive notification initiated without delay.

**7.3 Response Procedures**

1. **Detection & Analysis:** Identify scope, confirm incident type, and notify the Security Team.
2. **Containment:** Isolate affected systems, accounts, and datasets.
3. **Eradication:** Remove malicious code, revoke credentials, and clean affected data.
4. **Recovery:** Restore validated datasets and models from secure backups.
5. **Post-Incident Review:** Document root causes, lessons learned, and control improvements.

**7.4 Communication and Disclosure**

* Notify Legal, Compliance, and Executive teams per escalation protocol.
* Follow regulatory disclosure timelines as required (e.g., GDPR 72-hour rule).
* Communicate promptly and transparently with affected clients or partners.

**8. Enforcement and Sanctions**

* Non-compliance may result in disciplinary action, access revocation, contract termination, or legal escalation.
* Vendors found in violation of this policy may have contracts immediately suspended or terminated.

**9. References and Framework Alignment**

* **NIST AI Risk Management Framework (AI RMF)**
* **ISO/IEC 27001 – Information Security Management Systems**
* **ISO/IEC 23894 – AI Risk Management**
* **SOC 2 – Trust Services Criteria (Security, Availability, Confidentiality)**
* **Zero Trust Architecture (ZTA) Principles**

**10. Appendices**

**10.1 Key Definitions and Acronyms**

Includes standardized definitions such as AI RMF, RBAC, DPO, VDI, EDR, and DSA.

**10.2 Version Control and Approval Record**

Captures document revisions, dates, and approvals for traceability.

**10.3 Regulatory and Framework Alignment Map**

Maps policy sections to control objectives under ISO 27001, NIST AI RMF, and SOC 2.

**10.4 Sample Incident Reporting Workflow**

Visual or textual outline showing escalation steps and response timelines.

**10.5 Policy Intent and Summary**

Reiterates the objective of embedding **security-by-design** and **resilience-by-default** into AI model training pipelines.