Securing the AI Data Pipeline

CISO Checklist

Mandos mandos.io

Step 1: Know Your Data, Your AI, and Your Risks

☐ Inventory Your AI Landscape:

- ☐ **Identify all AI initiatives:** Create a comprehensive inventory of all AI projects and applications within your organization.

  Example
  This includes in-house developed AI models, third-party AI services, and even simple AI-powered features within existing software.

- ☐ **Categorize AI applications:** Classify each AI initiative based on its type and purpose.

  Example
  Is it a generative AI tool like a chatbot or content generator? A predictive model used for risk assessment? Understanding the specific capabilities and limitations of each AI application is key to assessing its potential risks.

- ☐ **Document data sources and flow:** For each AI initiative, identify the sources of data it uses, the types of data processed (e.g., customer data, financial information, proprietary code), and how that data flows through different systems.

  Example
  Does the AI application access data stored in cloud storage, on-premises databases, or external APIs? Visualizing these data flows using diagrams can help pinpoint potential vulnerabilities.

☐ Determine Your AI Scope:

- ☐ **Utilize the Generative AI Security Scoping Matrix:** This matrix, introduced in the sources, provides a framework for classifying AI usage into five scopes based on the level of control your organization has over the AI model and associated data.

Example
Using a public AI service like ChatGPT falls under Scope 1, while building and training a custom AI model from scratch corresponds to Scope 5. Each scope has a different risk profile, and understanding yours is crucial for tailoring your security approach.

▢ Map Your Data Flows:

- ▢ **Create visual representations:** Go beyond simple documentation and develop visual data flow diagrams for each AI application.
- ▢ **Highlight potential vulnerabilities:** Pinpoint areas where sensitive data is transmitted, stored, or processed. These are prime candidates for implementing security controls.

Example
If your data flow diagram reveals that an AI chatbot accesses a database containing customer PII, you'll need to ensure robust access controls and encryption measures are in place for that database.

Step 2: Build Your AI Security (Brick by Brick)

▢ Identity and Access Management (IAM):

- ▢ **Principle of Least Privilege:** This principle is paramount. Grant users and systems only the minimal access permissions necessary to perform their tasks. Overly permissive access is a recipe for disaster.
- ▢ **Strong Authentication:** Implement robust authentication mechanisms, including multi-factor authentication (MFA), to verify user identities before granting access to AI systems and data.
- ▢ **Fine-Grained Authorization:** Don't just rely on basic authentication. Utilize role-based access control (RBAC) to define granular permissions for different user roles. This ensures that users can only perform actions and access data relevant to their job function.
- ▢ **Control Access to Model Inference Endpoints:** Restrict direct access to the APIs that allow interaction with your AI models. Just like you wouldn't want unauthorized users connecting directly to your production database, you need to safeguard your AI models from unauthorized inference requests.

Example
If you're using Amazon Bedrock, leverage AWS Identity and Access Management (IAM) to manage permissions for invoking model inference.

- ▢ **Secure API Keys:** If your AI applications rely on API keys for authentication, treat these keys as highly sensitive information. Store them securely, rotate them regularly, and implement mechanisms to revoke compromised keys.

Data Protection, Always:

- **Encrypt Data at Rest and in Transit:** Encryption is your best friend. Encrypt sensitive data stored in databases, cloud storage, and any other location. Equally important is encrypting data in transit, especially when transmitting information over networks or to and from AI applications.

  Example
  If you're using Amazon S3 to store AI model artifacts, enable encryption for your S3 buckets. For sensitive data used in fine-tuning models, consider using a customer-managed AWS KMS key for an additional layer of control.

- **Data Minimization:** Don't collect or store data you don't absolutely need for your AI initiatives. The less data you have, the smaller your attack surface.

  Example
  If your AI model only needs to process customer names and email addresses, don't store sensitive information like social security numbers or credit card details.

- **De-Identification:** When possible, remove or mask personally identifiable information (PII) from your datasets. This is especially crucial for data used in training AI models, as it reduces the risk of exposing sensitive information.

Threat Modeling for the AI Age:

- **Understand AI-Specific Threats:** Traditional threat modeling techniques need to evolve to address the unique vulnerabilities of AI systems.

  Example
  Prompt injection, backdooring attacks, and data poisoning are just a few examples of threats specifically targeting AI applications.

- **Address Prompt Injection:** Prompt injection is a critical threat where attackers manipulate AI outputs by crafting malicious inputs. Mitigate this risk by:
  - **Sanitizing User Inputs:** Implement input validation and sanitization techniques to filter out potentially malicious characters or commands.
  - **Limiting Model Access to Sensitive Information:** Design your application to prevent the AI model from accessing or revealing information it shouldn't. The sources provide an example of redacting sensitive information like a customer's fraud score before it's included in the context provided to the model.
  - **Using Tools Like Amazon Bedrock Guardrails:** Bedrock Guardrails allow you to define policies and filters to detect and block undesirable content in both user inputs and AI outputs.
- **Secure the Supply Chain:** If you're using pre-trained AI models or third-party AI services, ensure the integrity and authenticity of the model artifacts.

Example
Verify the source of the model, use checksums or digital signatures to detect tampering, and scan models for known vulnerabilities.

Step 3: Empower Your People, Foster Collaboration, and Never Stop Learning

▢ Make Security Everyone's Business:

- ▢ **Regular Security Awareness Training:** Conduct ongoing security awareness training tailored to AI-related risks and best practices. Educate employees about:
  - ▢ **Data Handling Policies:** Clearly communicate policies regarding the handling of sensitive data, including data used in AI training and data generated by AI applications.
  - ▢ **AI-Specific Threats:** Explain the risks associated with AI, such as prompt injection and data poisoning, and provide examples of how these attacks can occur.
  - ▢ **Reporting Procedures:** Establish clear channels for reporting suspicious activity, potential data breaches, or concerns about AI security.
- ▢ **Develop a Culture of Security:** Encourage a security-first mindset where employees feel empowered to identify and report potential issues without fear of reprisal.

▢ Embrace the Power of Community:

- ▢ **Industry Collaboration:** Participate in industry forums, attend conferences, and engage with peers to share information about emerging AI security threats, best practices, and lessons learned.
- ▢ **Stay Informed:** The AI security landscape is constantly evolving. Subscribe to relevant security blogs, newsletters, and threat intelligence feeds to stay abreast of the latest developments.

▢ Continuous Monitoring and Improvement:

- ▢ **Implement Monitoring Tools:** Deploy security information and event management (SIEM) systems, intrusion detection and prevention systems (IDPS), and other security monitoring tools to detect and respond to suspicious activity within your AI infrastructure.
- ▢ **Regular Security Audits:** Conduct periodic security audits and penetration testing exercises to identify vulnerabilities in your AI systems and data pipelines.
- ▢ **Adapt and Improve:** Use the insights gained from monitoring, audits, and incident response to continuously improve your AI security posture.