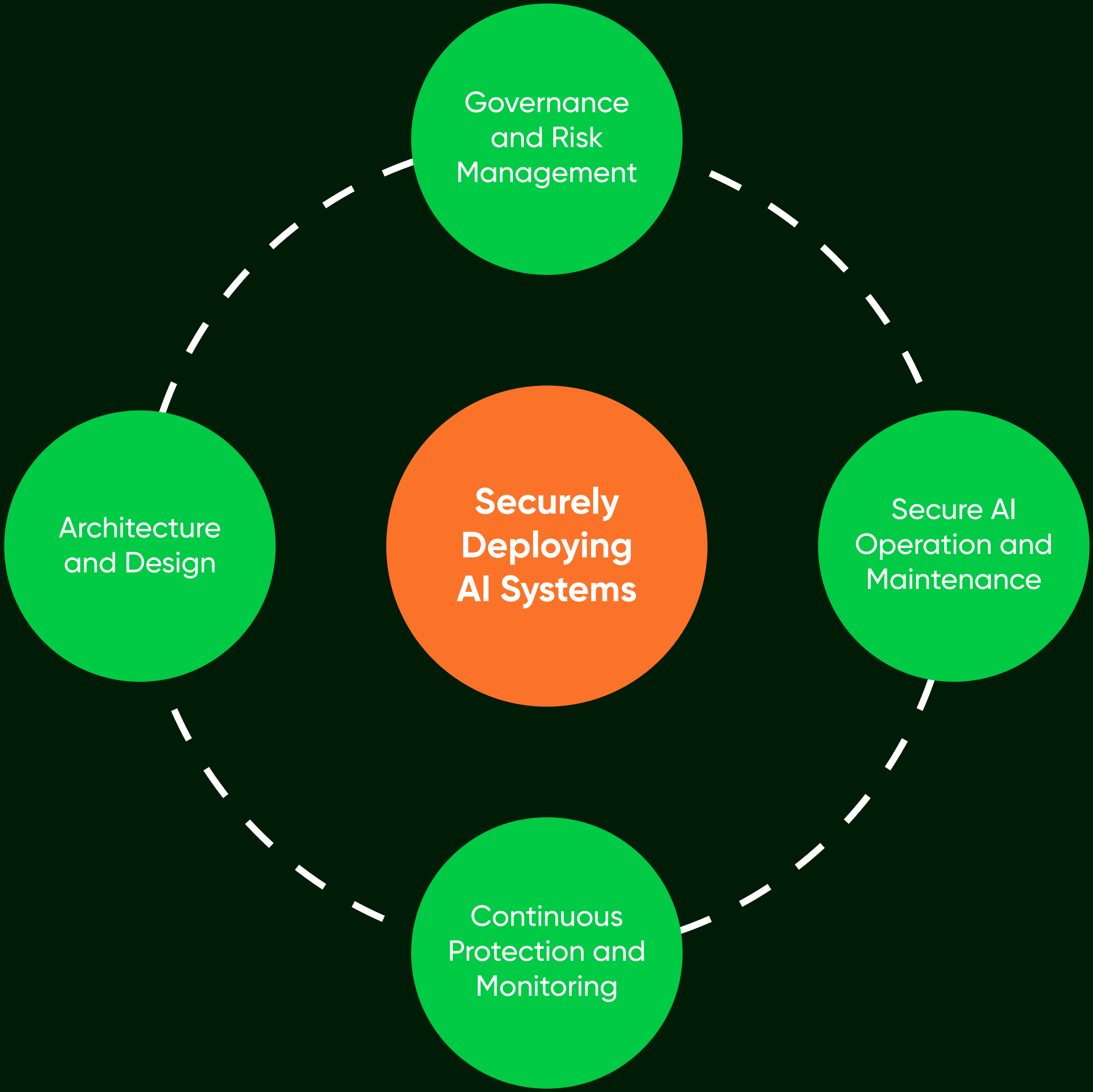
Comprehensive Checklist for Securely Deploying AI Systems

WITH CONFIDENCE STAVELEY

Sourced from the Joint Cybersecurity Information report on deploying AI Systems Security, these best practices provide a thorough and detailed checklist for deploying secure and resilient AI systems. Adhering to these guidelines will help your organization ensure AI systems are implemented securely and maintained with high cybersecurity standards.

NEXT PAGE >>>

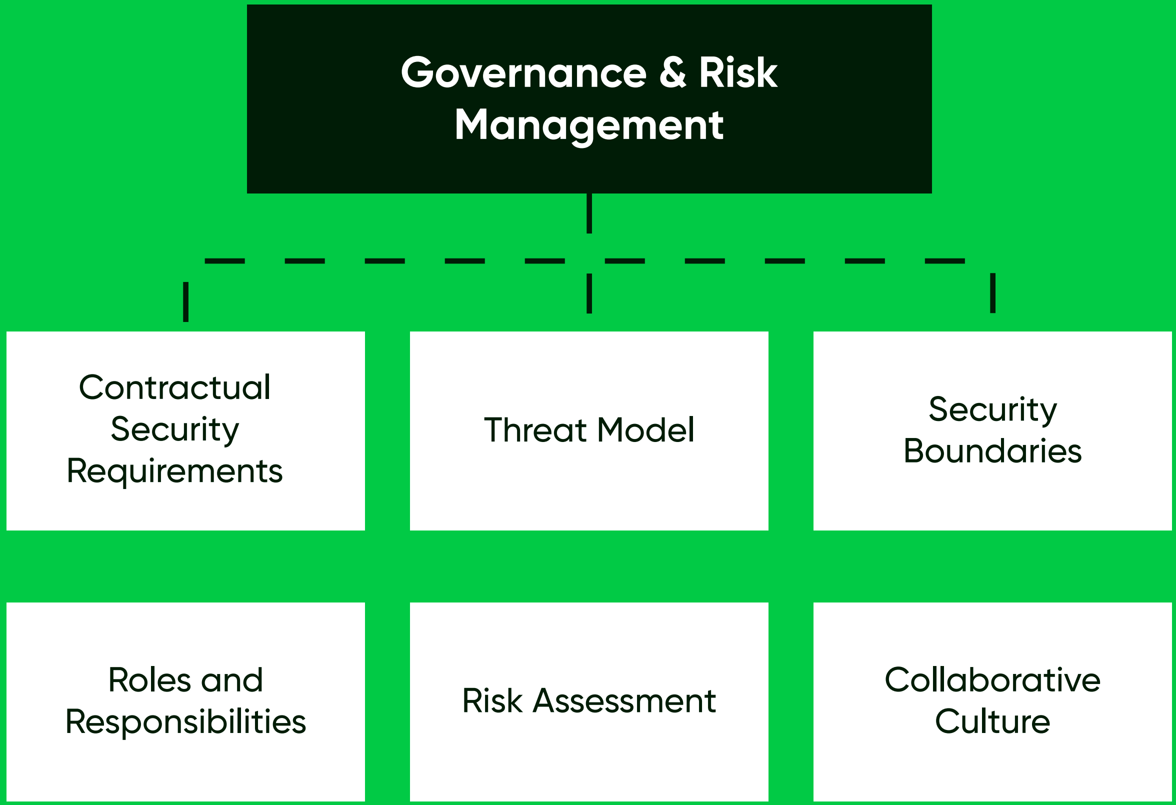

Governance and Risk Management
Architecture and Design
Securely Deploying AI Systems
Secure AI Operation and Maintenance
Continuous Protection and Monitoring

Confidence Staveley

NEXT PAGE >>>

1

Governance and Risk Management



Governance & Risk Management

Contractual Security Requirements

Threat Model

Security Boundaries

Roles and Responsibilities

Risk Assessment

Collaborative Culture

Confidence Staveley

NEXT PAGE >>>

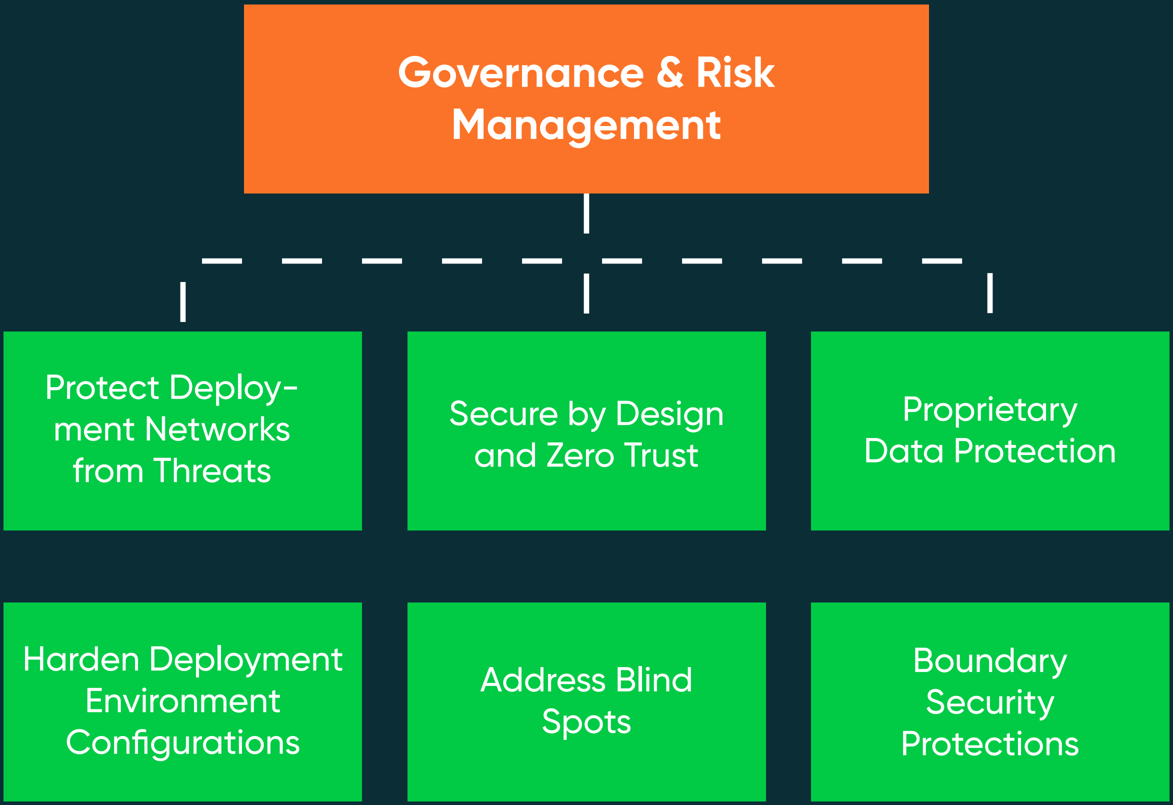
- Identify and confirm that the deployment environment meets stringent organizational IT standards.
- Facilitate coordination between AI and IT departments to ensure seamless integration.
- Conduct comprehensive threat assessments, documenting potential impacts and the organization’s risk tolerance.
- Implement a rigorous risk management framework to address identified vulnerabilities.
- Clearly delineate and document the roles, responsibilities, and accountability of all stakeholders.
- Ensure a robust chain of accountability for all security-related tasks.
- Promote open communication channels to address and mitigate any identified risks promptly.

NEXT PAGE >>>

- Define and enforce clear security boundaries within the IT environment for the AI system.
- Regularly review and update these boundaries to adapt to evolving threats.
- Obtain and meticulously analyze a detailed threat model from the AI system's primary developer.
- Leverage the threat model to implement proactive security measures and mitigation strategies.
- Incorporate stringent deployment environment security requirements into all AI system contracts.
- Ensure that these contracts mandate regular security audits and compliance checks.
- Foster a culture of collaboration among data science, infrastructure, and cybersecurity teams.

2

Architecture and Design



Governance & Risk Management

Protect Deployment Networks from Threats

Secure by Design and Zero Trust

Proprietary Data Protection

Harden Deployment Environment Configurations

Address Blind Spots

Boundary Security Protections

NEXT PAGE >>>

- Establish multi-layered security protections for the boundaries between the IT environment and the AI system.
- Implement advanced access control mechanisms to safeguard these boundaries.
- Conduct regular assessments to identify and address blind spots in boundary protections.
- Utilize access control systems for AI model weights, incorporating two-person control and integrity measures.
- Identify, classify, and protect all proprietary data sources used in AI model training and fine-tuning.
- Maintain a comprehensive catalog of trusted data sources to mitigate the risk of data poisoning or backdoor attacks.
- Apply secure-by-design principles to the architecture.
- Implement Zero Trust (ZT) frameworks to mitigate risks both to and from the AI system.

NEXT PAGE >>>

- Implement sandboxing for environments running ML models within hardened containers or VMs.
- Use advanced network monitoring and firewall configurations with allow lists.
- Stay abreast of hardware vendor guidance and notifications, applying software patches and updates via the Common Security Advisory Framework (CSAF).
- Encrypt AI model weights, outputs, and logs, storing encryption keys in a hardware security module (HSM).
- Implement robust authentication mechanisms and secure communication protocols.
- Use the latest version of Transport Layer Security (TLS) to encrypt data in transit.
- Ensure the use of phishing-resistant multifactor authentication (MFA) for access.

Confidence Staveley

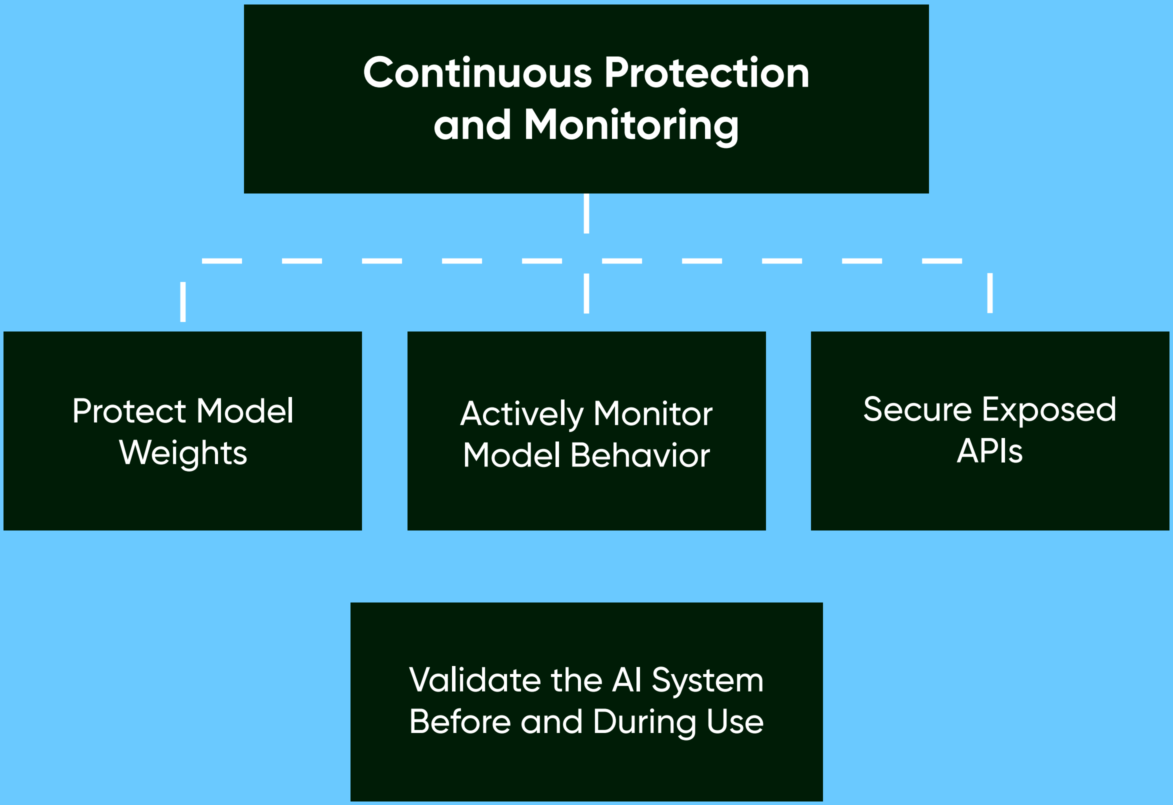
NEXT PAGE >>>

- Follow best practices for mitigating vulnerabilities as outlined in "Weak Security Controls and Practices Routinely Exploited for Initial Access."
- Adopt a Zero Trust security posture, assuming a breach is either inevitable or has already occurred.
- Implement advanced detection and response capabilities to enable rapid identification and containment of breaches.
- Utilize high-performing, industry-leading cybersecurity solutions for detecting unauthorized access.
- Integrate sophisticated incident detection systems to prioritize and respond to incidents with agility.

NEXT PAGE >>>

3

Continuous Protection and Monitoring



Continuous Protection and Monitoring

Protect Model Weights

Actively Monitor Model Behavior

Secure Exposed APIs

Validate the AI System Before and During Use

Confidence Staveley

NEXT PAGE >>>

- Utilize digital signatures and checksums to verify the origin and integrity of AI system artifacts.
- Encrypt safetensors to protect their integrity and confidentiality.
- Create and store hashes and encrypted copies of AI models in secure, tamper-proof locations.
- Store hash values and encryption keys in a secure vault or HSM to prevent unauthorized access.
- Employ version control systems with stringent access controls for all code and artifacts.
- Conduct rigorous testing of AI models for robustness, accuracy, and potential vulnerabilities.
- Apply adversarial testing to evaluate the model's resilience against sophisticated compromise attempts.

NEXT PAGE >>>

- Prepare for automated rollbacks with human-in-the-loop as a failsafe to ensure reliability and efficiency.
- Secure the supply chain for external AI models and data, adhering to the highest organizational standards.
- Prefer models developed according to secure-by-design principles.
- Inspect models within a secure development zone before tuning, training, and deployment.
- Use organization-approved AI-specific scanners for validating model integrity.
- Automate detection, analysis, and response capabilities to enhance IT and security team efficiency.
- Perform continuous scans of AI models and hosting IT environments to detect potential tampering.

NEXT PAGE >>>

- Implement robust authentication and authorization mechanisms for API access.
- Use secure protocols such as HTTPS with encryption and authentication.
- Implement rigorous validation and sanitization protocols for all input data to mitigate the risk of malicious input.
- Collect comprehensive logs for inputs, outputs, intermediate states, and errors.
- Automate alerts and triggers to facilitate real-time monitoring.
- Continuously monitor model architecture and configuration for unauthorized changes or modifications.
- Detect and respond to attempts to access or extract data from the AI model.

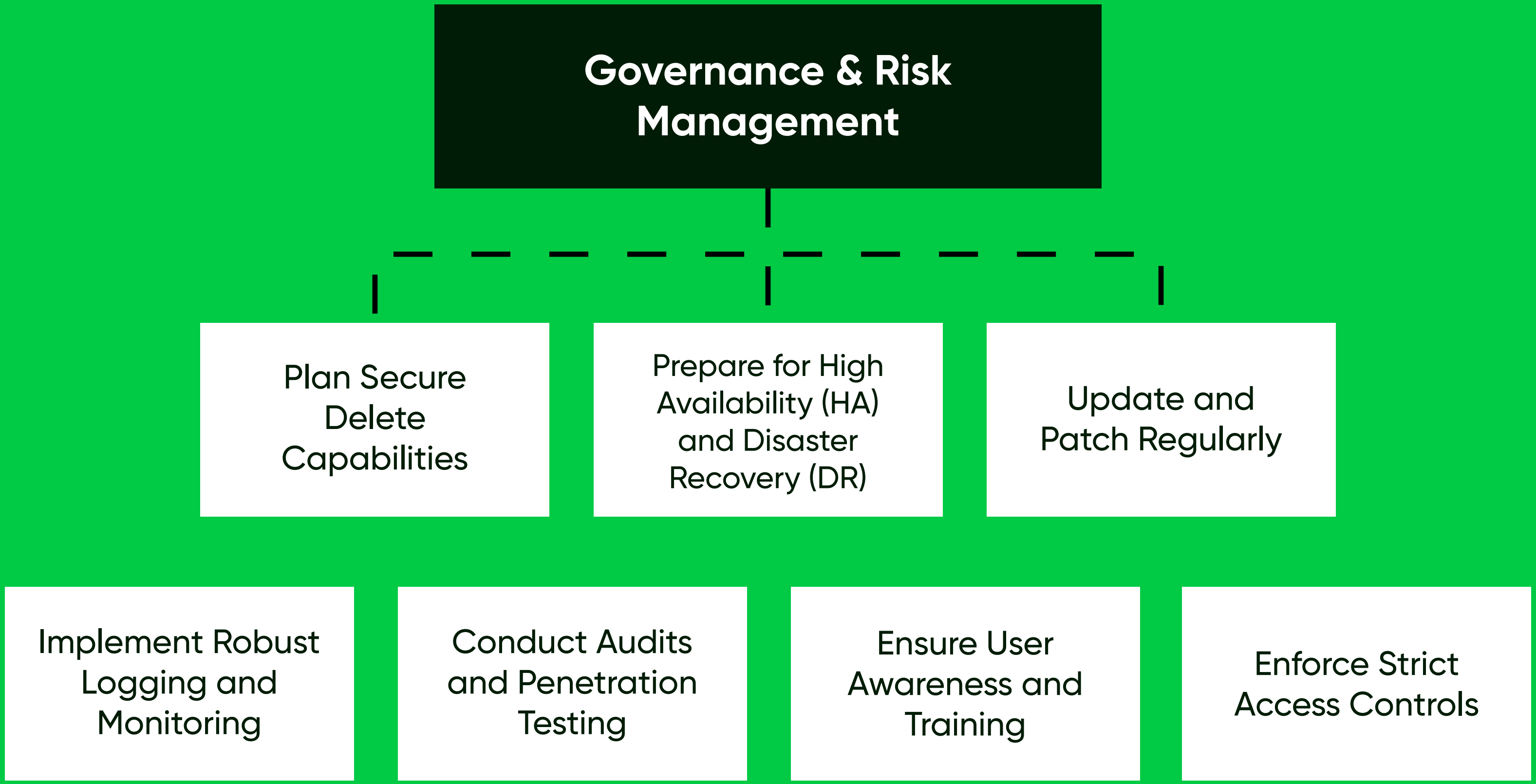
NEXT PAGE >>>

- Harden interfaces for accessing model weights to significantly increase the effort required for exfiltration.
- Ensure APIs return only the minimal data necessary for tasks to inhibit model inversion attacks.
- Implement advanced hardware protections for model weight storage, including disabling unnecessary communication capabilities.
- Aggressively isolate weight storage in protected storage vaults or highly restricted zones (HRZ), using hardware security modules (HSMs).

NEXT PAGE >>>

4

Secure AI Operation and Maintenance



Governance & Risk Management

Plan Secure Delete Capabilities

Prepare for High Availability (HA) and Disaster Recovery (DR)

Update and Patch Regularly

Implement Robust Logging and Monitoring

Conduct Audits and Penetration Testing

Ensure User Awareness and Training

Enforce Strict Access Controls

Confidence Staveley

NEXT PAGE >>>

- Apply robust role-based or attribute-based access controls to limit access to authorized personnel only.
- Require multifactor authentication (MFA) and privileged access workstations for administrative access.
- Educate users, administrators, and developers about advanced security best practices.
- Promote a pervasive security-aware culture to minimize the risk of human error.
- Utilize credential management systems to limit, manage, and monitor credential use rigorously.
- Engage with top-tier external security experts to conduct thorough audits and penetration testing of AI systems.
- Regularly update and refine security practices based on audit findings.

NEXT PAGE >>>

- Implement comprehensive monitoring of system behavior, inputs, and outputs.
- Establish sophisticated alert systems for potential security breaches or anomalies.
- Conduct full evaluations of AI models before deploying new versions to ensure security and performance standards are met.
- Use state-of-the-art immutable backup storage systems to ensure data and log immutability.
- Implement autonomous and irretrievable deletion of components post-process completion to ensure no residual data exposure.

Did you find this insightful?

Follow me for more content like this.

CONFIDENCE STAVELEY

Confidence Staveley