SECURING YOUR AI:

A STEP-BY-STEP GUIDE FOR CISOS

HIDDENLAYER

TABLE OF CONTENTS

02 **Foreword**

03 **Part 1: How Well Do You Know Your AI Environment**

Step 1: Establishing a Security Foundation
Step 2: Discovery and Asset Management
Step 3: Risk Assessment and Threat Modeling
Conclusion

08 **Part 2: Governing Your AI Systems**

Step 4: Defensive Frameworks
Step 5: Governance and Compliance
Step 6: Ethical AI Guidelines
Conclusion

12 **Part 3: Strengthen Your AI Systems**

Step 7: Data Security and Privacy
Step 8: Model Strength and Validation
Step 9: Secure Development Practices
Step 10: Continuous Monitoring and Incident Response
Step 11: Model Explainability and Transparency
Conclusion

19 **Part 4: Audit and Stay Up-to-Date on Your AI Environments**

Step 12: User Training and Awareness
Step 13: Third-Party Audits and Assessments
Step 14: Data Integrity and Quality
Step 15: Security Metrics and Reporting
Step 16: AI System Lifecycle Management
Step 17: Red Teaming Training
Step 18: Collaboration and Information Sharing
Conclusion: Securing Your AI Systems Effectively

FOREWORD

In today's digital landscape, the integration of Artificial Intelligence (AI) into business operations presents both unparalleled opportunities and unparalleled security challenges. As organizations race to adopt AI-driven solutions to enhance efficiency, customer experience, and decision-making processes, Chief Information Security Officers (CISOs) find themselves at the crossroads of innovation and risk management.

AI systems, with their ability to learn from vast datasets and adapt autonomously, have the potential to revolutionize every business process in any organization and touch every aspect of our society. Malicious actors are increasingly targeting AI systems to exploit weaknesses in data integrity, model behavior, and decision-making processes. Adversarial attacks, model poisoning, and AI manipulation are just a few examples of threats that CISOs must address to secure their organizations and protect their customers from harm.

The role of the CISO has never been more crucial in ensuring that AI deployments are both comprehensive and secure. If left unprotected, AI systems will become high-value targets for cyberattacks, leading to compromised data, eroded trust, and significant financial losses. Moreover, the ethical considerations surrounding AI—such as bias in decision-making algorithms, data privacy, and regulatory compliance—further complicate the security landscape. CISOs must navigate these challenges while fostering a culture of innovation and ensuring that AI systems are developed, deployed, and maintained with security as a foundational pillar.

This guide aims to equip CISOs with the knowledge, frameworks, and best practices to secure AI systems effectively. From understanding the unique risks posed by AI to implementing comprehensive defense strategies, this resource will help security leaders build resilient AI infrastructures. Whether your organization is in the early stages of AI adoption or has mature AI-driven systems, this guide will serve as a roadmap for mitigating risks and ensuring AI's safe and responsible use.

As you delve into the following pages, remember that securing AI is not just about technology—it's about trust. The confidence in the decisions made by AI, the integrity of the data it relies upon, and the resilience of the systems that support it are all critical to the future of cybersecurity. By proactively addressing the challenges posed by AI, CISOs can not only protect their organizations but also unlock the true potential of AI in a secure and ethical manner.

Malcolm Harkins
Chief Security and Trust Officer
at HiddenLayer

PART 01

HOW WELL DO YOU KNOW YOUR AI ENVIRONMENT

INTRODUCTION

As AI advances at a rapid pace, implementing comprehensive security measures becomes increasingly crucial. The integration of AI into critical business operations and society is growing, highlighting the importance of proactive security strategies. While there are concerns and challenges surrounding AI, there is also significant potential for leaders to make informed, strategic decisions. Organizational leaders can effectively navigate the complexities of AI security by seeking clear, actionable guidance and staying informed amidst the abundance of information. This proactive approach will help mitigate risks and ensure AI technologies' safe and responsible deployment, ultimately fostering trust and innovation.

Many existing frameworks and policies provide high-level guidelines but lack detailed, step-by-step instructions for security leaders. That's why we created "Securing Your AI: A Step-by-Step Guide for CISOs." This guide aims to fill that gap, offering clear, practical steps to help leaders worldwide secure their AI systems and dispel myths that can lead to insecure implementations.

Step

01 Establishing a Security Foundation

Establishing a strong security foundation is essential when beginning the journey to securing your AI. This involves understanding the basic principles of security for AI, setting up a dedicated security team, and ensuring all stakeholders know the importance of securing AI systems.

To begin this guide, we recommend reading our 2024 AI Threat Landscape Report, which covers the basics of securing AI.

2024 AI Threat Landscape Report

We also recommend the following persons to be involved and complete this step since they will be responsible for the following:

- **Chief Information Security Officer (CISO):** To lead the establishment of the security foundation.
- **Chief Information Officer (CIO) & Chief Technology Officer (CTO):** To provide strategic direction and resources.
- **AI Development Team:** To understand and integrate security principles into AI projects.
- **Compliance and Legal Team:** Ensure all security practices align with legal and regulatory requirements.

Ensuring these prerequisites are met sets the stage for successfully implementing the subsequent steps in securing your AI systems.

Now, let's begin.

Step

02 Discovery and Asset Management

Begin your journey by thoroughly understanding your AI ecosystem. This starts with conducting an AI usage inventory. Catalog every AI application and AI-enabled feature within your organization. For each tool, identify its purpose, origin, and operational status. This comprehensive inventory should include details such as:

- **Purpose**
  - What specific function does each AI application serve?
  - Is it used for data analysis, customer service, predictive maintenance, or another purpose?
- **Origin**
  - Where did the AI tool come from?
  - Was it developed in-house, sourced from a third-party vendor, or derived from an open-source repository?
- **Monitoring Netrwork Traffic**
  - Continuously monitor network traffic for unauthorized downloads of pre-trained models. This helps prevent rogue elements from infiltrating your system.

This foundational step is crucial for identifying potential vulnerabilities and gaps in your security infrastructure. By knowing exactly what AI tools are in use, you can better assess and manage their security risks.

Next, perform a pre-trained model audit. Track all pre-trained AI models sourced from public repositories. This involves:

- **Cataloging Pretrained Models:** Document all pre-trained models in use, noting their source, version, and specific use case within your organization.
- **Assessing Model Integrity:** Verify the authenticity and integrity of pre-trained models to ensure they have not been tampered with or corrupted.
- **Monitoring Network Traffic:** Continuously monitor network traffic for unauthorized downloads of pre-trained models. This helps prevent rogue elements from infiltrating your system.

Monitoring network traffic is essential to prevent unauthorized access and the use of pre-trained models, which can introduce security vulnerabilities. This vigilant oversight protects against unseen threats and ensures compliance with intellectual property and licensing agreements. Unauthorized use of pre-trained models can lead to legal and financial repercussions, so it is important to ensure that all models are used in accordance with their licensing terms.

By thoroughly understanding your AI ecosystem through an AI usage inventory and pre-trained model audit, you establish a strong foundation for securing your AI infrastructure. This proactive approach helps identify and mitigate risks, ensuring the safe and effective use of AI within your organization.

Who Should Be Responsible and In the Room:

- **Chief Information Security Officer (CISO):** To oversee the security aspects and ensure alignment with the overall security strategy.
- **Chief Technology Officer (CTO):** To provide insight into the technological landscape and ensure integration with existing technologies.
- **AI Team Leads (Data Scientists, AI Engineers):** To offer detailed knowledge about AI applications and models in use.
- **IT Managers:** To ensure accurate inventory and auditing of AI assets.
- **Compliance Officers:** To ensure all activities comply with relevant laws and regulations.
- **Third-Party Security Consultants:** If necessary, to provide an external perspective and expertise.

Step

03 Risk Assessment and Threat Modeling

With a clear inventory in place, assess the scope of your AI development. Understand the extent of your AI projects, including the number of dedicated personnel, such as data scientists and engineers, and the scale of ongoing initiatives. This assessment provides a comprehensive view of your AI landscape, highlighting areas that may require additional security measures. Specifically, consider the following aspects:

- **Team Composition:** Identify the number and roles of personnel involved in AI development. This includes data scientists, machine learning engineers, software developers, and project managers. Understanding your team structure helps assess resource allocation and identify potential skill gaps.
- **Project Scope:** Evaluate the scale and complexity of your AI projects. Are they small-scale pilots, or are they large-scale deployments across multiple departments? Assessing the scope helps understand the potential impact and the level of security needed.
- **Resource Allocation:** Determine the resources dedicated to AI projects, including budget, infrastructure, and tools. This helps identify whether additional investments are needed to bolster security measures.

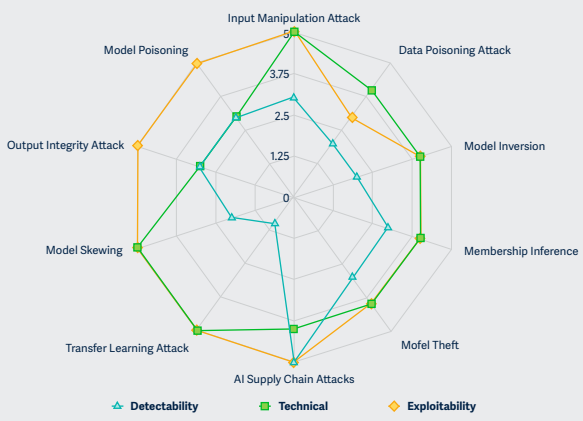
Afterward, a thorough risk and benefit analysis will be conducted. Identify and evaluate potential threats, such as data breaches, adversarial attacks, and misuse of AI systems. Simultaneously, assess the benefits to understand the value of these systems to your organization. This dual analysis helps prioritize security investments and develop strategies to mitigate identified risks effectively. Consider the following steps:

- **Risk Identification:** List all potential threats to your AI systems. These include data breaches, unauthorized access, adversarial attacks, model theft, and algorithmic bias. Consider both internal and external threats.
- **Risk Evaluation:** Assess the likelihood and impact of each identified risk. Determine how each risk could affect your organization in terms of financial loss, reputational damage, operational disruption, and legal implications.
- **Benefit Assessment:** Evaluate the benefits of your AI systems. This includes improved efficiency, cost savings, enhanced decision-making, competitive advantage, and innovation. Quantify these benefits to understand their value to your organization.
- **Prioritization:** Based on the risk and benefit analysis, prioritize your security investments. Focus on mitigating high-impact and high-likelihood risks first. Ensure that the benefits of your AI systems justify the costs and efforts of implementing security measures.

OSWASP ML SECURITY TOP 10

Security Risks	Exploitability	Detectability	Technical
Input Manipulation Attack	5	3	5
Data Poisoning Attack	3	2	4
Model Inversion	4	2	4
Membership Inference	4	3	4
Model Theft	4	3	4
AI Supply Chain Attacks	5	5	4
Transfer Learning Attack	5	1	5
Model Skewing	5	2	5
Output Integrity Attack	5	3	3
Model Poisoning	5	3	3

OSWASP ML TOP 10



By assessing the scope of your AI development and conducting a thorough risk and benefit analysis, you gain a holistic understanding of your AI landscape. This allows you to make informed decisions about where to allocate resources and how to mitigate risks effectively, ensuring the security and success of your AI initiatives.

Who Should Be Responsible and In the Room:

- **Risk Management Team:** To identify and evaluate potential threats.
- **Data Protection Officers:** To assess risks related to data breaches and privacy issues.
- **AI Ethics Board:** To evaluate ethical implications and misuse scenarios.
- **AI Team Leads (Data Scientists, AI Engineers):** To provide insights on technical vulnerabilities and potential adversarial attacks.
- **Business Analysts:** To understand and quantify these AI systems' benefits and value to the organization.
- **Compliance Officers:** To ensure all risk assessments are aligned with legal and regulatory requirements.
- **External Security Consultants:** To provide an independent assessment and validate internal findings.

PART 01

CONCLUSION

The first part of our guide has highlighted the often neglected importance of security for AI amidst the pressure from organizational leaders and the prevalence of misinformation. Organizations can begin their journey toward a secure AI ecosystem by establishing a strong security foundation and engaging key stakeholders. Organizations can identify potential vulnerabilities and establish a solid understanding of their AI assets, starting with a comprehensive AI usage inventory and pre-trained model audit. Moving forward, conducting a detailed risk assessment and threat modeling exercise will help prioritize security measures, aligning them with the organization's strategic goals and resources.

Through these initial steps, leaders can set the stage for a secure, ethical, and compliant AI environment, fostering trust and enabling the safe integration of AI into critical business operations. This proactive approach addresses current security challenges and prepares organizations to adapt to future advancements and threats in the AI landscape.

PART 02

GOVERNING YOUR AI SYSTEMS

INTRODUCTION

Effective governance ensures that AI systems are secure, ethical, and compliant with regulatory standards. As organizations increasingly rely on AI, they must adopt comprehensive governance strategies to manage risks, adhere to legal requirements, and uphold ethical principles. This second section focuses on the importance of defensive frameworks within a broader governance strategy. We explore how leading organizations have developed detailed frameworks to enhance security for AI and guide the development of ethical AI guidelines, ensuring responsible and transparent AI operations. Tune in as we continue to cover understanding AI environments, governing AI systems, strengthening AI systems, and staying up-to-date on AI developments over the next few weeks.

Step

04 Defensive Frameworks

As tools and techniques for attacking AI become more sophisticated, a methodical defensive approach is essential to safeguard AI. Over the past two years, leading organizations have developed comprehensive frameworks to enhance security for AI. Familiarizing yourself with these frameworks is crucial as you build out your secure AI processes and procedures. The following frameworks provide valuable guidance for organizations aiming to safeguard their AI systems against evolving threats.

MITRE ATLAS

MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) is a comprehensive framework launched in 2021, detailing adversarial machine learning tactics, techniques, and case studies. It complements the MITRE ATT&CK framework and includes real-world attacks and red-teaming exercises to provide a complete picture of AI system vulnerabilities.

In 2023, MITRE ATLAS was significantly updated, adding 12 new techniques and 5 unique case studies, focusing on large language models (LLMs) and generative AI systems. Collaborations with Microsoft led to new tools like the Arsenal and Almanac plugins for enhanced threat emulation. The update also introduced 20 new mitigations based on case studies. ATLAS now includes 14 tactics, 82 techniques, 22 case studies, and 20 mitigations, with ongoing efforts to expand its resources. This community-driven approach ensures that ATLAS remains a critical resource for securing AI-enabled systems against evolving threats.

NIST AI Risk Management Framework

Released in January 2023, the NIST AI RMF provides a conceptual framework for responsibly designing, developing, deploying, and using AI systems. It focuses on risk management through four functions: govern, map, measure, and manage.

Google Secure AI Framework (SAIF)

Introduced in June 2023, SAIF offers guidance on securing AI systems by adapting best practices from traditional software development. It emphasizes six core elements: expanding security foundations, automating defenses, and contextualizing AI risks.

OWASP Top 10

In 2023, OWASP released the Top 10 Machine Learning Risks, highlighting critical security risks in machine learning and providing guidance on prevention. Additionally, OWASP outlined vulnerabilities in large language models (LLMs), offering practical security measures.

Gartner AI Trust, Risk, and Security Management (AI TRiSM)

Gartner's AI TRiSM framework addresses bias, privacy, explainability, and security in AI/ML systems, providing a roadmap for building trusted, reliable, and secure AI systems.

Databricks AI Security Framework (DAISF)

Released in February 2024, DAISF provides a comprehensive strategy to mitigate cyber risks in AI systems, with actionable recommendations across 12 components of AI systems.

IBM Framework for Securing Generative AI

IBM's framework, released in January 2024, focuses on securing LLMs and generative AI solutions through five steps: securing data, models, usage, infrastructure, and establishing governance.

Step

05 Governance and Compliance

Ensuring compliance with relevant laws and regulations is the first step in creating ethical AI guidelines. Your AI systems must adhere to all legal and regulatory requirements, such as GDPR, CCPA, and industry-specific standards. Compliance forms the backbone of your security for AI strategy, helping you avoid legal pitfalls.

Who Should Be Responsible and In the Room:

- **Compliance and Legal Team:** Ensures AI systems meet all relevant laws and regulations, providing legal guidance and support.
- **Chief Information Security Officer (CISO):** Oversees the integration of compliance requirements into the overall security strategy.
- **AI Development Team:** Integrates compliance requirements into the design and development of AI systems.
- **Data Privacy Officer (DPO):** Ensures data protection practices comply with privacy laws such as GDPR and CCPA.
- **Chief Information Officer (CIO) & Chief Technology Officer (CTO)**: Provides oversight, resources, and strategic direction for compliance efforts.

Step

06 Ethical AI Guidelines

While working on Step 03: Risk Assessment and Threat Modeling, implement ethical AI guidelines to steer AI development and usage responsibly and transparently.

Start by forming an ethics committee that includes AI developers, data scientists, legal experts, ethicists, cybersecurity professionals, and, if needed, community representatives. This diverse group will oversee the creation and enforcement of the guidelines.

Identify core ethical principles such as fairness, transparency, accountability, privacy, and safety. Fairness ensures AI systems avoid biases and treat all users equitably. Transparency makes AI processes and decisions understandable to users and stakeholders. Accountability establishes clear lines of responsibility for AI outcomes. Privacy involves protecting user data through strong security measures and respecting user consent. Safety ensures AI systems operate securely and do not cause harm.

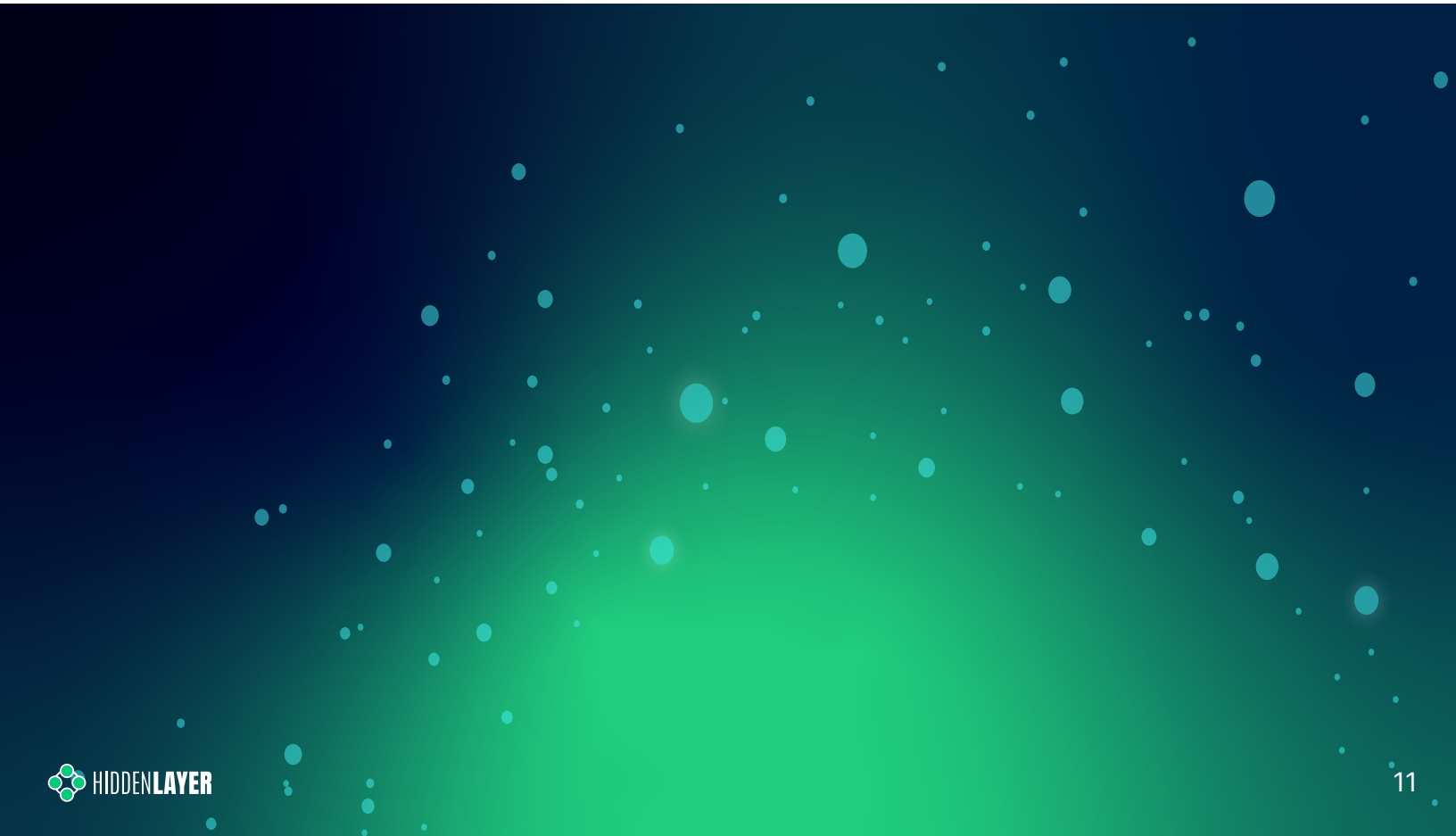
Consult internal and external stakeholders, including employees and customers, to gather insights. Draft the guidelines with a clear introduction, core ethical values, and specific measures for bias mitigation, data privacy, transparency, accountability, and safety. Circulate the draft for review, incorporating feedback to ensure the guidelines are comprehensive and practical.

Once finalized, conduct training sessions for all employees involved in AI development and deployment. Make the guidelines accessible and embed ethical considerations into every stage of the AI lifecycle. Establish a governance framework for ongoing oversight and regular audits to ensure compliance and address emerging ethical issues. Regularly update the guidelines to reflect new insights and encourage continuous feedback from stakeholders.

PART 02

CONCLUSION

Effective governance is essential for managing AI systems in an era of sophisticated threats and stringent regulatory requirements. By integrating comprehensive defensive frameworks such as MITRE ATLAS, NIST AI RMF, Google SAIF, OWASP Top 10, Gartner AI TRiSM, Databricks AI Security Framework, and IBM's generative AI framework, organizations can enhance the security of their AI systems. However, governance goes beyond security; it encompasses ensuring compliance with laws and regulations, such as GDPR and CCPA, and embedding ethical principles into AI development and deployment. Forming a diverse ethics committee and establishing clear guidelines on fairness, transparency, accountability, privacy, and safety are crucial steps in this process. By embedding these principles into every stage of the AI lifecycle and maintaining ongoing oversight, organizations can build and sustain AI systems that are not only secure but also ethical and trustworthy.



PART 03

STRENGTHEN YOUR AI SYSTEMS

INTRODUCTION

Strengthening your AI systems is crucial to ensuring their security, reliability, and trustworthiness. This third section focuses on implementing advanced measures to secure data, validate models, embed secure development practices, monitor systems continuously, and ensure model explainability and transparency.

These steps are essential for protecting sensitive information, maintaining user trust, and complying with regulatory standards.

Step

07 Data Security and Privacy

Data is the lifeblood of AI. Deploy advanced security measures tailored to your AI solutions that are adaptable to various deployment environments. This includes implementing encryption, access controls, and anonymization techniques to protect sensitive data. Ensuring data privacy is critical in maintaining user trust and complying with regulations.

Evaluate third-party vendors rigorously. Your vendors must meet stringent security for AI standards. Integrate their security measures into your overall strategy to ensure there are no weak links in your defense. Conduct thorough security assessments and require vendors to comply with your security policies and standards.

Who Should Be Responsible and In the Room:

- **Data Security Team:** Implements encryption, access controls, and anonymization techniques.
- **AI Development Team:** Ensures AI solutions are designed with integrated security measures.
- **Compliance and Legal Team:** Ensures compliance with data privacy regulations.
- **Third-Party Vendor Management Team:** Evaluates and integrates third-party vendor security measures.
- **Chief Information Officer (CIO) & Chief Technology Officer (CTO):** Provides oversight and resources for security initiatives.

Step

08 Model Strength and Validation

AI models must be resilient to ensure their reliability and effectiveness. Regularly subject them to adversarial testing to evaluate their systems. This process involves simulating various attacks to identify potential vulnerabilities and assess the model's ability to withstand malicious inputs. By doing so, you can pinpoint weaknesses and fortify the model against potential threats.

Employing thorough model validation techniques is equally essential. These techniques ensure consistent, reliable behavior in real-world scenarios. For example, cross-validation helps verify that the model performs well across different subsets of data, preventing overfitting and ensuring generalizability. Stress testing pushes the model to its limits under extreme conditions, revealing how it handles unexpected inputs or high-load situations.

Both adversarial testing and validation processes are critical for maintaining trust and reliability in your AI outputs. They provide a comprehensive assessment of the model's performance, ensuring it can handle the complexities and challenges of real-world applications. By integrating these practices into your AI development and maintenance workflows, you can build more resilient and trustworthy AI systems.

Who Should Be Responsible and In the Room:

- **AI Development Team:** Designs and develops the AI models, ensuring strength and the ability to handle adversarial testing.
- **Data Scientists:** Conduct detailed analysis and validation of the AI models, including cross-validation and stress testing.
- **Cybersecurity Experts:** Simulate attacks and identify vulnerabilities to test the model's resilience against malicious inputs.
- **Quality Assurance (QA) Team:** Ensures that the AI models meet required standards and perform reliably under various conditions.
- **Chief Information Officer (CIO) & Chief Technology Officer (CTO):** Provides oversight, resources, and strategic direction for testing and validation processes.

ADVERSARIAL TESTING METHODS

Feature Extraction Testing	**Feature Extraction Fuzzing:** Finding bugs in the feature extractor by subjecting it to strange input
Model Testing	**Time-Based Validation:** If your data has a time component in its collection, train the model on historical data and validate on newer data **Out-Of-Distribution Testing:** Attempt to find samples that are quite different from samples existing in your training data to see how well your model generalizes to these samples **Feature Importance:** Check the impact individual features can have on the overall model classification or output. if there are features that attackers can easily control that have an outsized influence on the model output, consider finding ways to remove the attacker's ability to control this feature

Step

09 Secure Development Practices

Embed security best practices at every stage of the AI development lifecycle. From inception to deployment, aim to minimize vulnerabilities by incorporating security measures at each step. Start with secure coding practices, ensuring that your code is free from common vulnerabilities and follows the latest security guidelines. Conduct regular code reviews to catch potential security issues early and to maintain high standards of code quality.

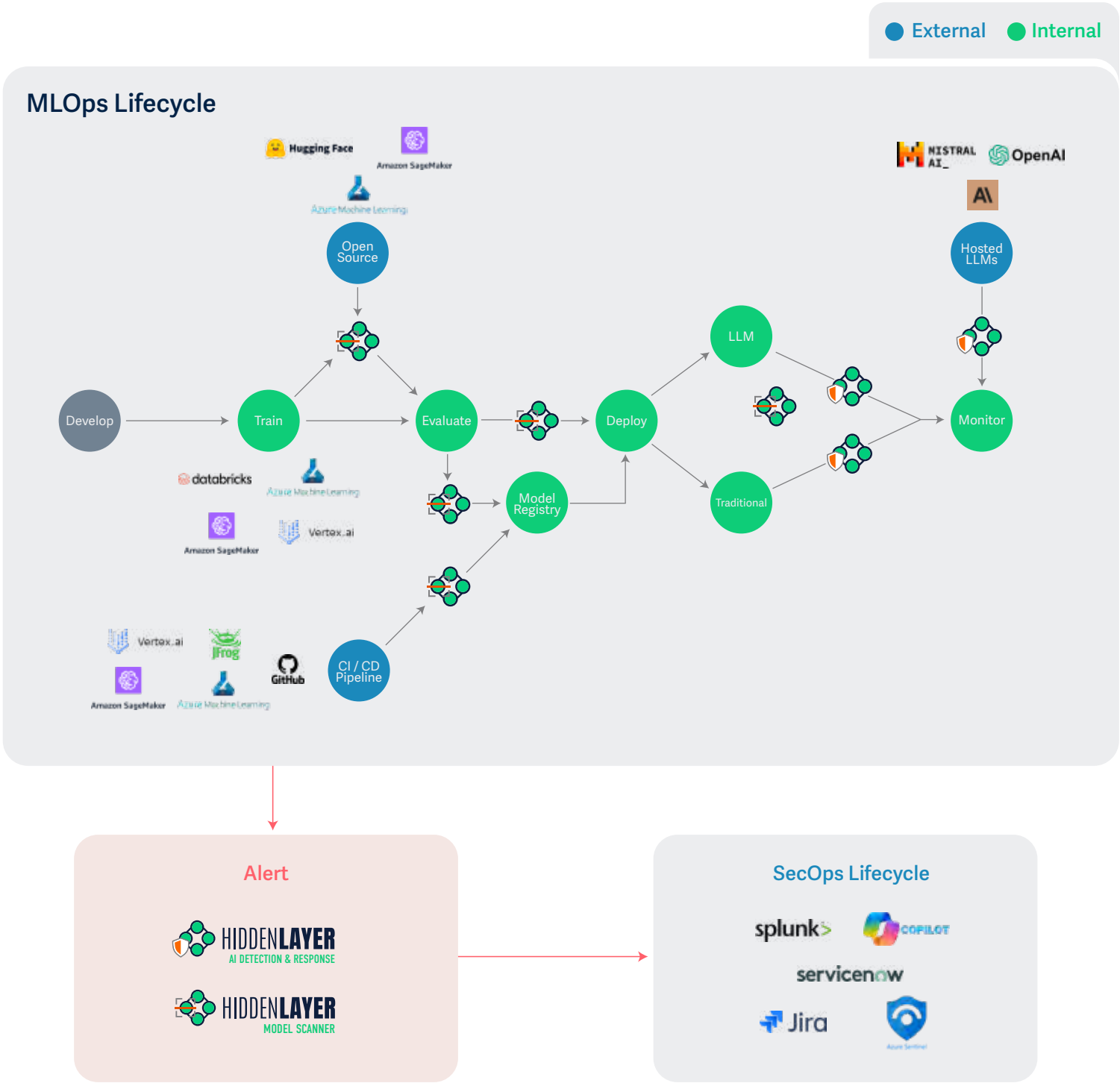
Implement comprehensive security testing throughout the development process. This includes static and dynamic code analysis, penetration testing, and vulnerability assessments. These tests help identify and mitigate risks before they become critical issues. Additionally, threat modeling should be incorporated to anticipate potential security threats and design defenses against them.

By embedding these secure development practices, you ensure that security is integrated into your AI systems from the ground up. This proactive approach significantly reduces the risk of introducing vulnerabilities during development, leading to strong and secure AI solutions. It also helps maintain user trust and compliance with regulatory requirements, as security is not an afterthought but a fundamental component of the development lifecycle.

Who Should Be Responsible and In the Room:

- **AI Development Team:** Responsible for secure coding practices and incorporating security measures into the AI models from the start.
- **Security Engineers:** Conduct regular code reviews, static and dynamic code analysis, and penetration testing to identify and address security vulnerabilities.
- **Cybersecurity Experts:** Perform threat modeling and vulnerability assessments to anticipate potential security threats and design appropriate defenses.
- **Quality Assurance (QA) Team:** Ensures that security testing is integrated into the development process and that security standards are maintained throughout.
- **Project Managers:** Coordinate efforts across teams, ensuring that security best practices are followed at every stage of the development lifecycle.
- **Compliance and Legal Team:** Ensures that the development process complies with relevant security regulations and industry standards.
- **Chief Information Officer (CIO) & Chief Technology Officer (CTO):** Provides oversight, resources, and support for embedding security practices into the development lifecycle.

One Platform For All Your Security for AI Needs



Step

10 Continuous Monitoring and Incident Response

Implement continuous monitoring systems to detect anomalies immediately to ensure the ongoing security and integrity of your AI systems. Real-time surveillance acts as an early warning system, enabling you to identify and address potential issues before they escalate into major problems. These monitoring systems should be designed to detect a wide range of indicators of compromise, such as unusual patterns in data or system behavior, unauthorized access attempts, and other signs of potential security breaches.

Advanced monitoring tools should employ machine learning algorithms and anomaly detection techniques to identify deviations from normal activity that may indicate a threat. These tools can analyze vast amounts of data in real time, providing comprehensive visibility into the system's operations and enabling swift response to any detected anomalies.

Additionally, integrating continuous monitoring with automated response mechanisms can further enhance security. When an anomaly is detected, automated systems can trigger predefined actions, such as alerting security personnel, isolating affected components, or initiating further investigation procedures. This proactive approach minimizes the time between detection and response, reducing the risk of significant damage.

To effectively implement continuous monitoring systems for immediately detecting anomalies, it's crucial to consider products specifically designed for this purpose. Involving the right stakeholders to evaluate and select these products ensures a strong and effective monitoring strategy.

Pair continuous monitoring with a comprehensive incident response strategy. Regularly update and rehearse this strategy to maintain readiness against evolving threats, as preparedness is key to effective incident management. An effective incident response plan includes predefined roles and responsibilities, communication protocols, and procedures for containing and mitigating incidents.

A Ponemon survey found that 77% of respondents lack a formal incident response plan that is consistently applied across their organization, and nearly half say their plan is informal or nonexistent. Don't be part of the 77% who do not have an up-to-date incident response (IR) plan. It's time for security to be proactive rather than reactive, especially regarding AI.

For support on developing an incident response plan, refer to the CISA guide on Incident Response Plan Basics. This guide provides valuable insights into what an IR plan should include and needs.

Step

11 Model Explainability and Transparency

Before starting Step 11, make sure you have fully completed Step 6: implement Ethical AI Guidelines.

As you know, transparency and explainability are critical, especially when it comes to improving the public's trust in AI usage. Ensure AI decisions can be interpreted and explained to users and stakeholders. Explainable AI builds trust and ensures accountability by making the decision-making process understandable. Techniques such as model interpretability tools, visualizations, and detailed documentation are essential for achieving this goal.

Regularly publish transparency reports detailing AI system operations and decisions. Transparency is not just about compliance; it's about fostering an environment of openness and trust. These reports should provide insights into how AI models function, the data they use, and the measures taken to ensure their fairness and reliability.

Who Should Be Responsible and In the Room:

- **AI Development Team:** Implements model interpretability tools, visualizations, and detailed documentation to make AI decisions interpretable and explainable.
- **Data Scientists:** Develop techniques and tools for explaining AI models and decisions, ensuring these explanations are accurate and accessible.
- **Compliance and Legal Team:** Ensures transparency practices comply with relevant regulations and industry standards, providing guidance on legal and ethical requirements.
- **Communication and Public Relations Team:** Publishes regular transparency reports and communicates AI system operations and decisions to users and stakeholders, fostering an environment of openness and trust.

PART 03

CONCLUSION

Strengthening your AI systems requires a multi-faceted approach encompassing data security, model validation, secure development practices, continuous monitoring, and transparency. Organizations can protect sensitive data and ensure compliance with privacy regulations by implementing advanced security measures such as encryption, access controls, and anonymization techniques. Rigorous evaluation of third-party vendors and adversarial testing of AI models further enhance the reliability and resilience of AI systems.

Embedding secure development practices throughout the AI lifecycle, from secure coding to regular security testing, helps minimize vulnerabilities and build strong, secure AI solutions. Continuous monitoring and a well-defined incident response plan ensure that potential threats are detected and addressed promptly, maintaining the integrity of AI systems. Finally, fostering transparency and explainability in AI decisions builds trust and accountability, making AI systems more understandable and trustworthy for users and stakeholders.

By following these comprehensive steps, organizations can create AI systems that are not only secure but also ethical and transparent, ensuring they serve as valuable and reliable assets in today's complex technological landscape

PART 04

AUDIT AND STAY UP-TO-DATE ON YOUR AI ENVIRONMENTS

INTRODUCTION

In this final section, we will explore essential topics for comprehensive AI systems: data security and privacy, model validation, secure development practices, continuous monitoring, and model explainability. Key areas include encryption, access controls, anonymization, and evaluating third-party vendors for security compliance. We will emphasize the importance of red teaming training, which simulates adversarial attacks to uncover vulnerabilities. Techniques for adversarial testing and model validation will be discussed to ensure AI robustness. Embedding security best practices throughout the AI development lifecycle and implementing continuous monitoring with a strong incident response strategy is crucial.

Step

12 User Training and Awareness

Continuous education is vital. Conduct regular training sessions for developers, data scientists, and IT staff on security best practices for AI. Training should cover topics such as secure coding, data protection, and threat detection. An informed team is your first line of defense against security threats.

Raise awareness across the organization about security for AI risks and mitigation strategies. Knowledge is power, and an aware team is a proactive team. Regular workshops, seminars, and awareness campaigns help keep security top of mind for all employees.

Who Should Be Responsible and In the Room:

- **Training and Development Team:** Organizes and conducts regular training sessions for developers, data scientists, and IT staff on security for AI best practices.
- **AI Development Team:** Participates in training on secure coding, data protection, and threat detection to stay updated on the latest security measures.
- **Data Scientists:** Engages in ongoing education to understand and implement data protection and threat detection techniques.
- **IT Staff:** Receives training on security for AI best practices to ensure strong implementation and maintenance of security measures.
- **Security Team:** Provides expertise and updates on the latest security for AI threats and mitigation strategies during training sessions and awareness campaigns.

Step

13 Third-Party Audits and Assessments

Engage third-party auditors to review your security for AI practices regularly. Fresh perspectives can identify overlooked vulnerabilities and provide unbiased assessments of your security posture. These auditors bring expertise from a wide range of industries and can offer valuable insights that internal teams might miss. Audits should cover all aspects of security for AI, including data protection, model robustness, access controls, and compliance with relevant regulations. A thorough audit assesses the entire lifecycle of AI deployment, from development and training to implementation and monitoring, ensuring comprehensive security coverage.

Conduct penetration testing on AI systems periodically to find and fix vulnerabilities before malicious actors exploit them. Penetration testing involves simulating attacks on your AI systems to identify weaknesses and improve your defenses. This process can uncover flaws in your infrastructure, applications, and algorithms that attackers could exploit. Regularly scheduled penetration tests, combined with ad-hoc testing when major changes are made to the system, ensure that your defenses are constantly evaluated and strengthened. This proactive approach helps ensure your AI systems remain resilient against emerging threats as new vulnerabilities are identified and addressed promptly.

In addition to penetration testing, consider incorporating other forms of security testing, such as red team exercises and vulnerability assessments, to provide a well-rounded understanding of your security posture. Red team exercises simulate real-world attacks to test the effectiveness of your security measures and response strategies.

Vulnerability assessments systematically review your systems to identify and prioritize security risks. Together, these practices create a strong security testing framework that enhances the resilience of your AI systems.

CYBERSECURITY FRAMEWORKS LIST

	Framework	Industry
01	**SOC 2**	Service providers such as data centers, SaaS companies, managed service providers, cloud computing providers
02	**ISO 27001**	Finance, healthcare, IT, government sectors
03	**NIST Framework**	Critical infrastructure sectors like energy, healthcare, finance, transportation
04	**HIPAA**	Healthcare providers, health plans, healthcare clearinghouses
05	**PCI DSS**	Merchants, financial institutions, payment processors
06	**GDPR**	Businesses, government agencies, non-profits
07	**HITRUST CSF**	Healthcare organizations and business associates
08	**COBIT**	Organizations of all sizes and industries
09	**NERC-CIP**	Electric utilities, power generation companies
10	**FISMA**	U.S. federal government agencies and contractors
11	**NIST Special Publication 800-53**	U.S. federal agencies and organizations
12	**NIST Special Publication 800-171**	Non-federal organizations handling controlled unclassified information for the U.S. government
13	**IAB CCPA**	Businesses collecting personal information from California residents
14	**CIS CONTROLS**	Organizations of all sizes and sectors
15	**UK Telecoms (Security) Act 2021**	Telecommunication companies operating in the United Kingdom
16	**CISA Telecoms Framework**	Telecom providers operating in the United States
17	**Cyber Essentials Plus**	UK companies and businesses in various industries
18	**FedRAMP**	U.S. federal agencies and cloud service providers serving government clients

By engaging third-party auditors and regularly conducting penetration testing, you improve your security for AI posture and demonstrate a commitment to maintaining high-security standards. This can enhance trust with stakeholders, including customers, partners, and regulators, by showing that you take proactive measures to protect sensitive data and ensure the integrity of your AI solutions.

Who Should Be Responsible and In the Room:

- **Chief Information Security Officer (CISO):** Oversees security for AI practices and the engagement with third-party auditors.
- **Security Operations Team:** Manages security audits and penetration testing, and implements remediation plans.
- **IT Security Manager:** Coordinates with third-party auditors and facilitates the audit process.
- **AI Development Team Lead:** Addresses vulnerabilities identified during audits and testing, ensuring strong AI model security.
- **Compliance Officer:** Ensures security practices comply with regulations and implements auditor recommendations.
- **Risk Management Officer:** Integrates audit and testing findings into the overall risk management strategy.
- **Chief Information Officer (CIO) & Chief Technology Officer (CTO):** Provides oversight, resources, and strategic direction for security initiatives.

Step

14 Data Integrity and Quality

Implement strong procedures to ensure the quality and integrity of data used for training AI models. Begin with data quality checks by establishing validation and cleaning processes to maintain accuracy and reliability.

Regularly audit your data to identify and fix any issues, ensuring ongoing integrity. Track the origin and history of your data to prevent using compromised or untrustworthy sources, verifying authenticity and integrity through data provenance.

Maintain detailed metadata about your datasets to provide contextual information, helping assess data reliability. Implement strict access controls to ensure only authorized personnel can modify data, protecting against unauthorized changes.

Document and ensure transparency in all processes related to data quality and provenance. Educate your team on the importance of these practices through training sessions and awareness programs.

Who Should Be Responsible and In the Room:

- **Data Quality Team:** Manages data validation and cleaning processes to maintain accuracy and reliability.
- **Audit and Compliance Team:** Conducts regular audits and ensures adherence to data quality standards and regulations.
- **Data Governance Officer:** Oversees data provenance and maintains detailed records of data origin and history.
- **IT Security Team:** Implements and manages strict access controls to protect data integrity.
- **AI Development Team:** Ensures data quality practices are integrated into AI model training and development.
- **Training and Development Team:** Educates staff on data quality and provenance procedures, ensuring ongoing awareness and adherence.

Step

15 Security Metrics and Reporting

Define and monitor key security metrics to gauge the effectiveness of your security for AI measures. Examples include the number of detected incidents, response times, and the effectiveness of security controls.

Review and update these metrics regularly to stay relevant to current threats. Benchmark against industry standards and set clear goals for continuous improvement. Implement automated tools for real-time monitoring and alerts.

Establish a clear process for reporting security incidents, ensuring timely and accurate responses. Incident reports should detail the nature of the incident, affected systems, and resolution steps. Train relevant personnel on these procedures.

Conduct root cause analysis for incidents to prevent future occurrences, building a resilient security framework. To maintain transparency and a proactive security culture, communicate metrics and incident reports regularly to all stakeholders, including executive leadership.

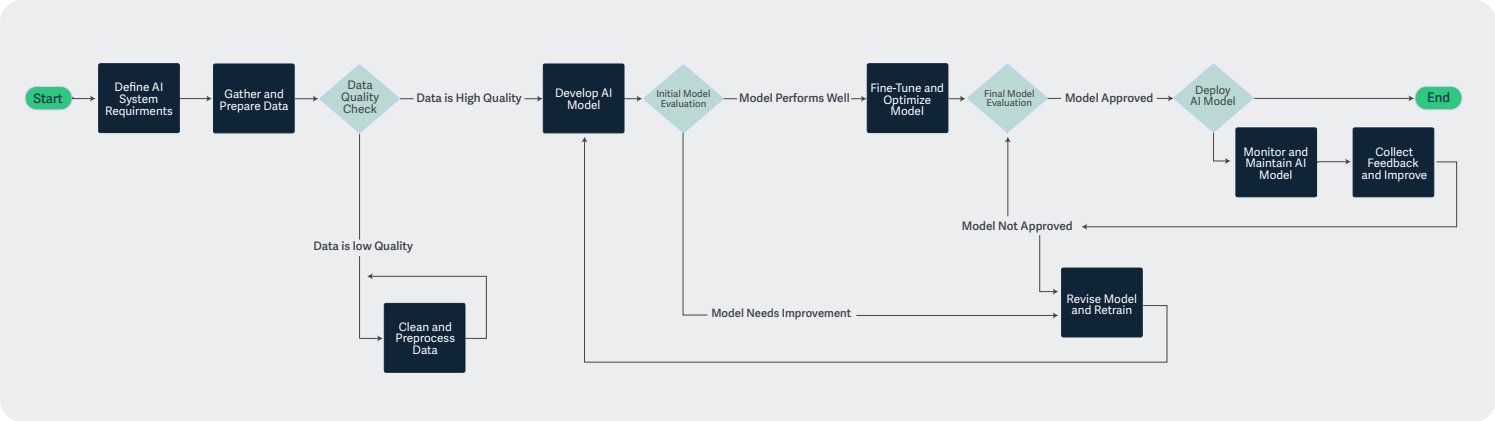
Who Should Be Responsible and In the Room:

- **Chief Information Security Officer (CISO):** Oversees the overall security strategy and ensures the relevance and effectiveness of security metrics.
- **Security Operations Team:** Monitors security metrics, implements automated tools, and manages real-time alerts.
- **Data Scientists:** Analyze security metrics data to provide insights and identify trends.
- **IT Security Manager:** Coordinates the reporting process and ensures timely and accurate incident reports.
- **Compliance and Legal Team:** Ensures all security measures and incident reports comply with relevant regulations.
- **Chief Information Officer (CIO) & Chief Technology Officer (CTO):** Reviews security metrics and incident reports to maintain transparency and support proactive security measures.

SAMPLE OF A SECURITY DASHBOARD

Step

16 AI System Lifecycle Management



Manage AI systems from development to decommissioning, ensuring security at every stage of their lifecycle. This comprehensive approach includes secure development practices, continuous monitoring, and proper decommissioning procedures to maintain security throughout their operational lifespan. Secure development practices involve implementing security measures from the outset, incorporating best practices in secure coding, data protection, and threat modeling. Continuous monitoring entails regularly overseeing AI systems to detect and respond to security threats promptly, using advanced monitoring tools to identify anomalies and potential vulnerabilities.

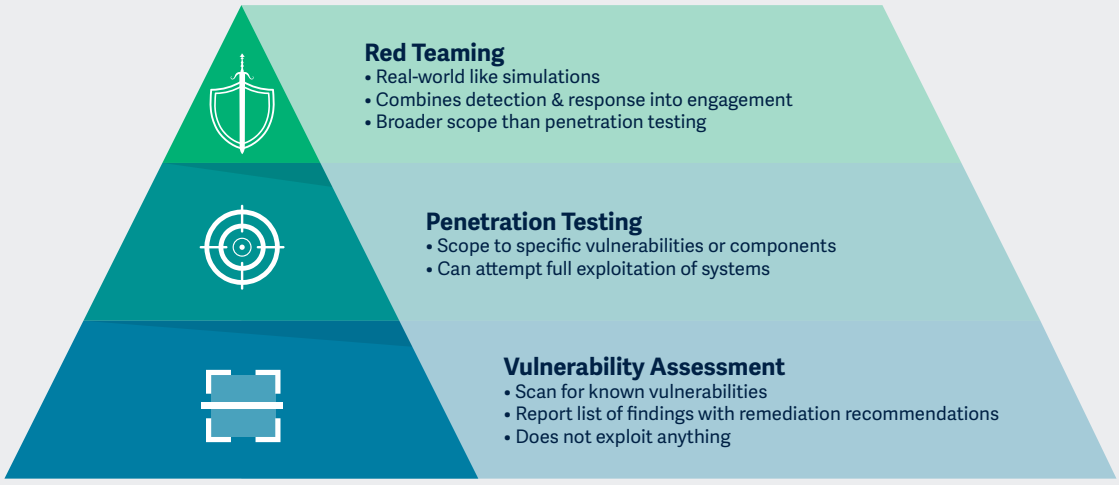
Proper decommissioning procedures are crucial when retiring AI systems. Follow stringent processes to securely dispose of data and dismantle infrastructure, preventing unauthorized access or data leaks. Clearly defining responsibilities ensures role clarity, making lifecycle management cohesive and strong. Effective communication is essential, as it fosters coordination among team members and strengthens your AI systems' overall security and reliability.

Who Should Be Responsible and In the Room:

- **Chief Information Security Officer (CISO):** Oversees the entire security strategy and ensures all stages of the AI lifecycle are secure.
- **AI Development Team:** Implements secure development practices and continuous monitoring.
- **IT Infrastructure Team:** Handles the secure decommissioning of AI systems and ensures proper data disposal.
- **Compliance and Legal Team:** Ensures all security practices meet legal and regulatory requirements.
- **Project Manager:** Coordinates efforts across teams, ensuring clear communication and role clarity.

Step

17 Red Teaming Training



To enhance the security of your AI systems, implement red teaming exercises. These involve simulating real-world attacks to identify vulnerabilities and test your security measures. If your organization lacks well-trained AI red teaming professionals, it is crucial to engage reputable external organizations, such as HiddenLayer, for specialized AI red teaming training to ensure comprehensive security.

To start the red teaming training, assemble a red team of cybersecurity professionals. Once again, given that your team may not be well-versed in security for AI enlist outside organizations to provide the necessary training. Develop realistic attack scenarios that mimic potential threats to your AI systems. Conduct these exercises in a controlled environment, closely monitor the team's actions, and document each person's strengths and weaknesses.

Analyze the findings from the training to identify knowledge gaps within your team and address them promptly. Use these insights to improve your incident response plan where necessary. Schedule quarterly red teaming exercises to test your team's progress and ensure continuous learning and improvement.

Integrating red teaming into your security strategy, supported by external training as needed, helps proactively identify and mitigate risks. This ensures your AI systems are robust, secure, and resilient against real-world threats.

Step

18 Collaboration and Information Sharing

Collaborate with industry peers to share knowledge about security for AI threats and best practices. Engaging in information-sharing platforms keeps you informed about emerging threats and industry trends, helping you stay ahead of potential risks. By collaborating, you can adopt best practices from across the industry and enhance your own security measures.

For further guidance, check out our blog post, From National Security to Building Trust: The Current State of Securing AI, which delves into the benefits of collaboration in securing AI. The blog provides valuable insights and practical advice on how to effectively engage with industry peers to strengthen your security for AI posture.

CONCLUSION:

Securing Your AI Systems Effectively

Securing AI systems is an ongoing, dynamic process that requires a thorough, multi-faceted approach. As AI becomes deeply integrated into the core operations of businesses and society, the importance of strong security measures cannot be overstated. This guide has provided a comprehensive, step-by-step approach to help organizational leaders navigate the complexities of securing AI, from initial discovery and risk assessment to continuous monitoring and collaboration.

By diligently following these steps, leaders can ensure their AI systems are secure but also trustworthy and compliant with regulatory standards. Implementing secure development practices, continuous monitoring, and rigorous audits, coupled with a strong focus on data integrity and collaboration, will significantly enhance the resilience of your AI infrastructure.

At HiddenLayer, we are here to guide and assist organizations in securing their AI systems. Don't hesitate to reach out for help. Our mission is to support you in navigating the complexities of securing AI ensuring your systems are safe, reliable, and compliant. We hope this series helps provide guidance on securing AI systems at your organization.

Remember: Stay informed, proactive, and committed to security best practices to protect your AI systems and, ultimately, your organization's future. For more detailed insights and practical advice, be sure to explore our blog post on collaboration in security for AI and our comprehensive Threat Report.