A Practitioner's Guide to Secure AI Implementation

Seven critical security concepts from "The CISO's Guide to AI-Powered Security" that every AI practitioner must address.

As AI is being streamlined in our daily lives. Whether it is to improve employee productivity, to increase the speed to perform a repetitive task or to improve security of an organization from modern day threats, every practitioner in security, SOC, IT operations and developer teams is thinking about AI.

1 Employees are using unsanctioned GenAI tools, creating "Shadow AI"

CHALLENGE:

This introduces massive risk by potentially feeding sensitive corporate data into unsanctioned, untrusted models without any security oversight or governance.

TAKEAWAY FOR PRACTITIONERS:

You cannot secure what you cannot see. Work with security teams to gain visibility into AI application usage. Advocate for sanctioned, secure tools and educate colleagues on the risks of using public AI with proprietary data.

2 Integrate existing applications with AI using APIs

CHALLENGE:

AI introduces new attack vectors across the entire lifecycle not just in the final application. Attack vectors exist in the training data (poisoning), the model itself (extraction), and the supply chain (vulnerable open-source components).

TAKEAWAY FOR PRACTITIONERS:

Think beyond securing just the final application. Your security posture must cover data sourcing, the training environment, third-party model integrity, and continuous testing for vulnerabilities.

3 Use open source and third-party components for faster development cycle

CHALLENGE:

Your AI is only as secure as its weakest open-source component. The data emphasizes the heavy reliance on open-source models, libraries, and datasets. Each element is a potential entry point for vulnerabilities, malicious code, or data with embedded biases.

TAKEAWAY FOR PRACTITIONERS:

Rigorously vet all third-party and open-source components. Use software composition analysis (SCA) tools and maintain a software bill of materials (SBOM) for your AI/ML projects. Don't implicitly trust pre-trained models.

4 Stay up to date with modern day threats

CHALLENGE:

Classic vulnerabilities have new AI-specific equivalents, like Prompt Injection and require new approaches. The OWASP Top 10 for Large Language Model Applications provides a formal framework for understanding that inputs can be weaponized to manipulate your model and bypass security filters.

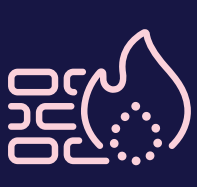
TAKEAWAY FOR PRACTITIONERS:

Familiarize yourself with the OWASP Top 10 for LLMs. Treat all user-provided inputs as potentially malicious and implement robust input validation and output encoding to mitigate these known risks.

5 Protect your AI / LLM / SLM models

CHALLENGE:

Existing network firewalls are not enough. Effective AI security requires implementing strong, centralized guardrails that provide real-time inline policies for data handling, acceptable use, and risk management that apply to all AI development and deployment.



TAKEAWAY FOR PRACTITIONERS:

Don't build in a vacuum. Engage with your organization's security and governance teams early. Adhere to established guardrails for data privacy, model testing, and security checks to ensure your innovations are built on a secure foundation.

6 Attackers are using GenAI to create more sophisticated and evasive threats

CHALLENGE:

Adversaries are now leveraging the same AI technology to author polymorphic malware that evades traditional detection and to craft hyper-realistic phishing attacks at an unprecedented scale.

TAKEAWAY FOR PRACTITIONERS:

The tools you are building with are also being used to create attacks against you. This elevates the need for zero trust security principles, as AI-powered threats can more easily bypass legacy defenses and manipulate employees. Your security awareness is more critical than ever.

7 Generate content to help improve employee productivity

CHALLENGE:

AI models can inadvertently generate content that infringes on copyrights or exposes proprietary data. There is a legal and reputational risk that a model, trained on vast and varied datasets, might reproduce copyrighted material or leak sensitive information it was trained on, creating significant liability for the organization.

TAKEAWAY FOR PRACTITIONERS:

Understand the provenance and licensing of your training data. Implement output filtering and monitoring to detect and block the generation of potentially infringing content or the regurgitation of sensitive training data before it reaches the end-user.

To Learn More:

As you are implementing AI in your environments, keep a close eye on these 7 concepts. To learn more about these, please visit us here.

LEARN MORE

Act Fast.
Stay Secure.