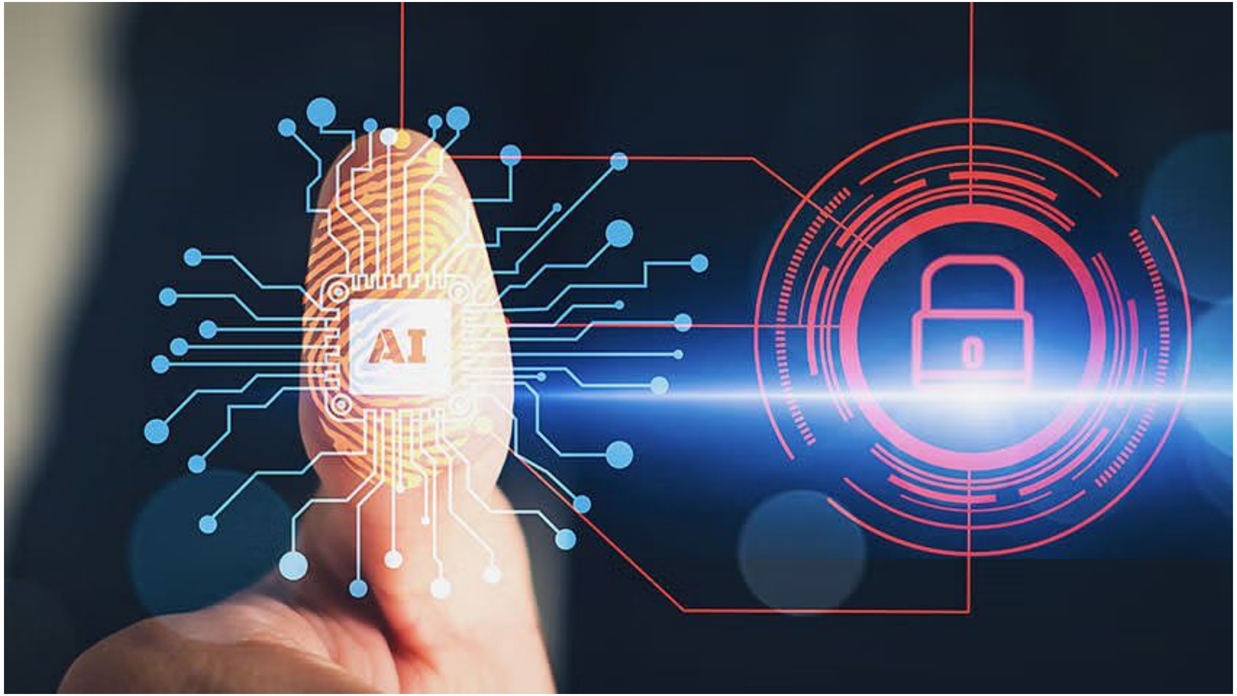
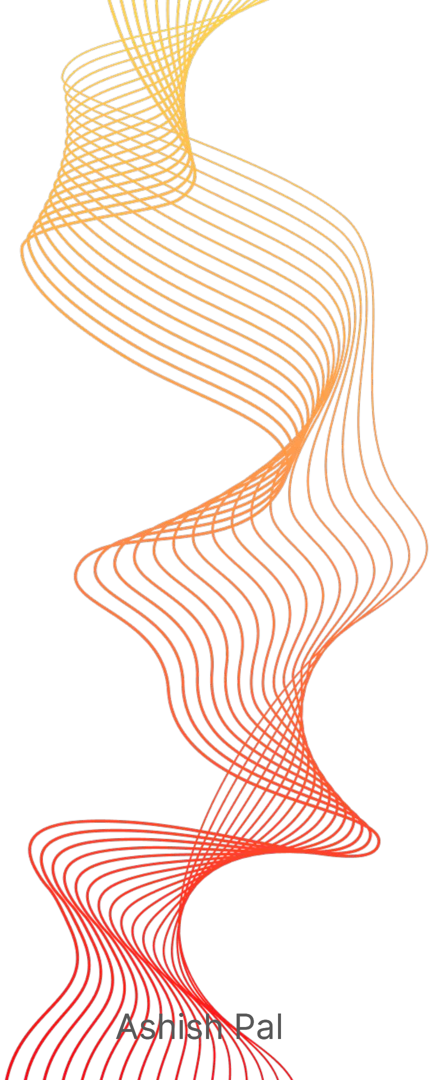
AI Privacy and Ethical



Ashish Pal

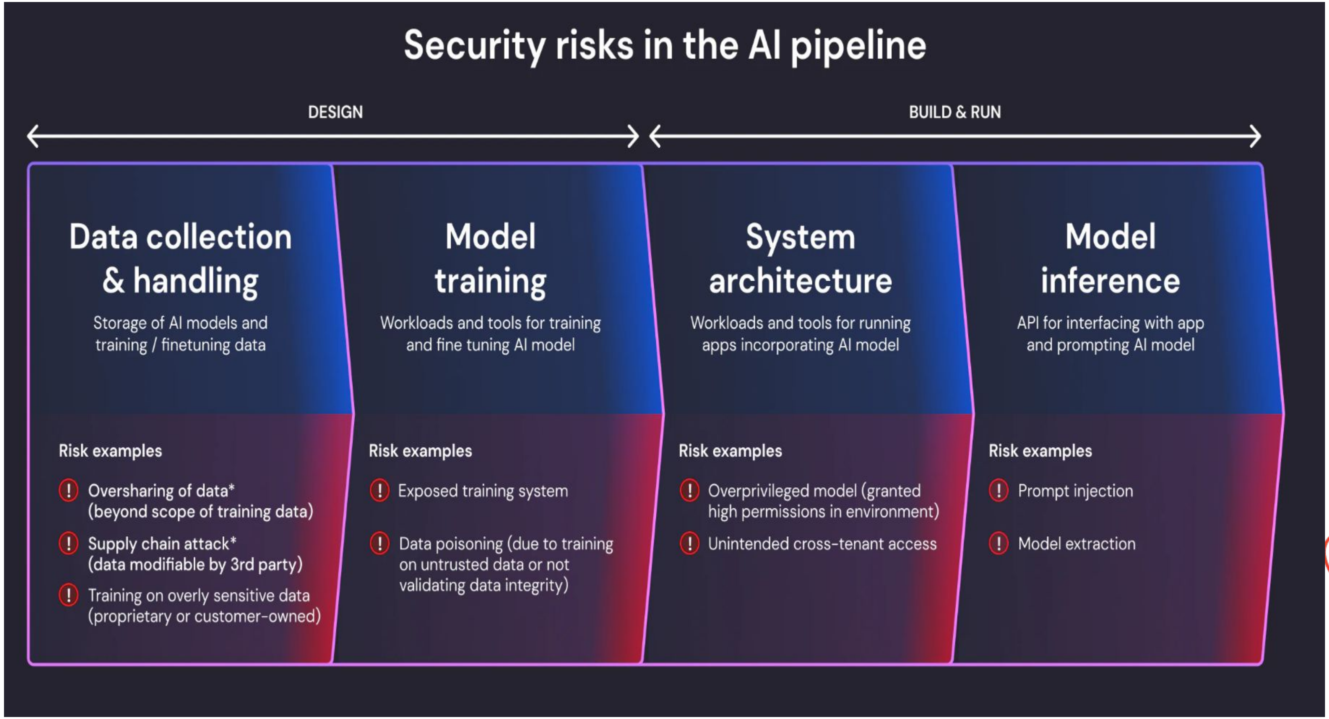
Index

1. Privacy risks in AI/ML applications
2. Understanding differential privacy
3. AI security frameworks and standards
4. Ethical considerations in AI Security
5. Hands-on Lab



Ashish Pal

Privacy risks in AI/ML applications



Security risks in the AI pipeline

DESIGN ⟷ BUILD & RUN

Data collection & handling

Storage of AI models and training / finetuning data

Risk examples

- ! Oversharing of data* (beyond scope of training data)
- ! Supply chain attack* (data modifiable by 3rd party)
- ! Training on overly sensitive data (proprietary or customer-owned)

Model training

Workloads and tools for training and fine tuning AI model

Risk examples

- ! Exposed training system
- ! Data poisoning (due to training on untrusted data or not validating data integrity)

System architecture

Workloads and tools for running apps incorporating AI model

Risk examples

- ! Overprivileged model (granted high permissions in environment)
- ! Unintended cross-tenant access

Model inference

API for interfacing with app and prompting AI model

Risk examples

- ! Prompt injection
- ! Model extraction

Ashish Pal

Understanding differential privacy

Blog

Ashish Pal

Ethical considerations in AI Security

Blog

Ethical AI Book

AI security frameworks and standards

- NIST’s Artificial Intelligence Risk Management framework breaks down AI security into four primary functions: govern, map, measure, and manage.
- Mitre’s Sensible Regulatory Framework for AI Security and ATLAS Matrix anatomize attack tactics and propose certain AI regulations.
- OWASP’s Top 10 for LLMs identifies and proposes standards to protect the most critical vulnerabilities associated with LLMs, such as prompt injections, supply chain vulnerabilities, and model theft.
- Google’s Secure AI Framework offers a six-step process to mitigate the challenges associated with AI systems. These include automated cybersecurity fortifications and risk-based management.
- PEACH framework emphasizes tenant isolation via privilege hardening, encryption hardening, authentication hardening, connectivity hardening, and hygiene (P.E.A.C.H.). Tenant isolation is a design principle that breaks down your cloud environments into granular segments with tight boundaries and stringent access controls.

Ashish Pal

A few simple AI security recommendations and best practices

1. Embrace an agile, cross-functional mindset

For most organizations, employees are already using AI technology, and AI use cases have been developed. The first draft of your AI framework needs to be developed fast to ensure a general security foundation for existing AI processes. Following an agile mindset, security teams should then define a priority mechanism that can further specialize the AI framework to your organization's AI requirements through short iterative update cycles.

In support of this evolving AI framework definition, establish a culture of open communication around AI security from the very beginning. Encouraging dialogue and collaboration ensures that potential risks are identified and mitigated efficiently while providing a way for security teams to communicate and enforce AI security requirements.

Ashish Pal

A few simple AI security recommendations and best practices

2. Understand the threat landscape for AI

For most organizations, employees are already using AI technology, and AI use cases have been developed. The first draft of your AI framework needs to be developed fast to ensure a general security foundation for existing AI processes. Following an agile mindset, security teams should then define a priority mechanism that can further specialize the AI framework to your organization’s AI requirements through short iterative update cycles.

In support of this evolving AI framework definition, establish a culture of open communication around AI security from the very beginning. Encouraging dialogue and collaboration ensures that potential risks are identified and mitigated efficiently while providing a way for security teams to communicate and enforce AI security requirements.

A few simple AI security recommendations and best practices

3. Define the AI security requirements for your organization

Different organizations have different security requirements, and no one-size-fits-all framework exists for AI security.

To fortify your security foundation, it’s imperative to establish comprehensive organization-centric governance policies. These policies should include a spectrum of considerations, ranging from data privacy and asset management to ethical guidelines and compliance standards. Since AI is a discipline driven by open-source contributions, third-party risk management is particularly relevant to ensure security for AI

To effectively manage AI-related risks, security teams should adopt a proactive stance where protocols are continuously evaluated and adapted. Security controls like ongoing system behavior monitoring, regular penetration testing, and the implementation of resilient incident response plans are indispensable.

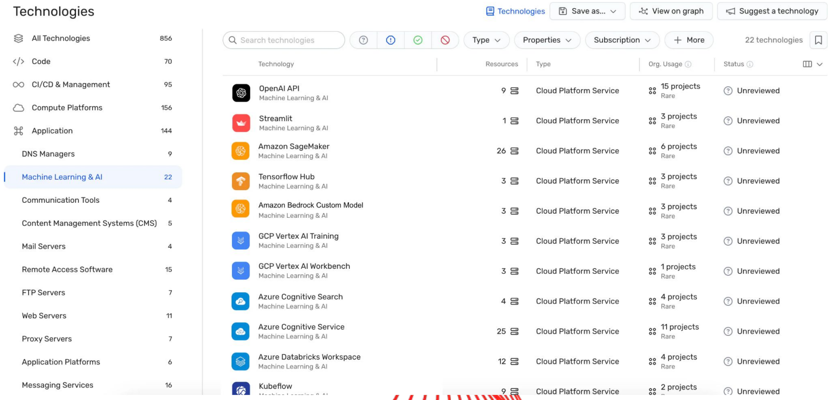
A few simple AI security recommendations and best practices

4. Ensure comprehensive visibility

Security can only be achieved for processes that are known and visible.

The first step for security teams seeking comprehensive visibility across all AI applications is to create and maintain an AI bill of materials (AI-BOM). An AI-BOM is an inventory of all components and dependencies within an organization's AI systems—whether in-house, third-party, or open-source.

Before accepting and introducing an AI application in the AI-BOM, it's a good idea to create a templated AI-model card. The AI-model card clearly and concisely documents all relevant details of the AI model for various stakeholders, including adherence to security requirements.



All Technologies	856
Code	70
CI/CD & Management	95
Compute Platforms	156
Application	144
DNS Managers	9
Machine Learning & AI	22
Communication Tools	4
Content Management Systems (CMS)	5
Mail Servers	4
Remote Access Software	15
FTP Servers	7
Web Servers	11
Proxy Servers	7
Application Platforms	6
Messaging Services	16

Technology	Resources	Type	Org. Usage	Status
OpenAI API Machine Learning & AI	9	Cloud Platform Service	15 projects Rare	Unreviewed
Streamlit Machine Learning & AI	1	Cloud Platform Service	3 projects Rare	Unreviewed
Amazon SageMaker Machine Learning & AI	26	Cloud Platform Service	6 projects Rare	Unreviewed
Tensorflow Hub Machine Learning & AI	3	Cloud Platform Service	3 projects Rare	Unreviewed
Amazon Bedrock Custom Model Machine Learning & AI	3	Cloud Platform Service	3 projects Rare	Unreviewed
GCP Vertex AI Training Machine Learning & AI	3	Cloud Platform Service	3 projects Rare	Unreviewed
GCP Vertex AI Workbench Machine Learning & AI	3	Cloud Platform Service	1 projects Rare	Unreviewed
Azure Cognitive Search Machine Learning & AI	4	Cloud Platform Service	4 projects Rare	Unreviewed
Azure Cognitive Service Machine Learning & AI	25	Cloud Platform Service	11 projects Rare	Unreviewed
Azure Databricks Workspace Machine Learning & AI	12	Cloud Platform Service	4 projects Rare	Unreviewed
Kubeflow Machine Learning & AI	9	Cloud Platform Service	2 projects	Unreviewed

Ashish Pal

A few simple AI security recommendations and best practices

5. Allow only safe models and vendors

As part of your AI framework, security teams should establish a rigorous vetting process to evaluate any external AI models and vendors against predefined security requirements. External AI solutions to be vetted include frameworks, libraries, model weights, and datasets. At a minimum, your security requirements should encompass data encryption and data handling, access control, and adherence to industry standards, including certifications. Any external AI solution that successfully passes this process is expected to be trustworthy and secure.

A few simple AI security recommendations and best practices

6. Implement automated security testing

Regularly scanning AI models and applications with specialized tools allows security teams to proactively identify vulnerabilities. These checks may include classic tests such as scanning for container security and dependencies or fuzz testing, as well as AI-specific scans via tools such as Alibi Detect or the Adversarial Robustness Toolbox. Make sure your teams test AI applications against misconfigurations or configuration mismatches, which could serve as easy entry points for security breaches. Being able to detect attack paths throughout the AI pipelines, from sensitive training data and exposed secrets to identities and network exposures, before they become threats in production is your goal.

Finally, functional testing is also a necessity. To ensure that the core functionalities of the AI applications are safe, or security implications are known and documented, functional testing includes classic unit testing and integration testing as well as AI-specific testing for data validation, model performance, model regression, and ethicality as supported by bias and fairness analysis.

A few simple AI security recommendations and best practices

7. Focus on continuous monitoring

Beyond testing, the dynamic and inherently non-deterministic nature of AI systems requires ongoing vigilance. Focus on continuous monitoring to sustain a secure and reliable AI ecosystem that can successfully address unexpected AI behavior and misuse.

Establish a robust system for monitoring both AI applications and infrastructure to detect anomalies and potential issues in real-time. Real-time monitoring processes track key performance indicators, model outputs, data distribution shifts, model performance fluctuations, and other system behaviors.

A few simple AI security recommendations and best practices

8. Raise staff awareness of threats and risks

As the AI framework for your organization matures in tandem with advancements in the field of SecOps for AI, security teams need to dedicate time to educating staff about threats and risks so that each individual AI user adheres to basic security guidelines.

First, it’s best practice for security teams to collaborate with data science teams to provide clear and concise security guidelines. The design of these security guidelines should promote experimentation for data science teams as much as possible. This way, you minimize the risk of data science teams neglecting or bypassing security controls to unlock the potential of AI.

After the first security guidelines are in place, you should offer comprehensive training to all employees to equip the entire workforce with the knowledge to use AI safely.

DataBricks

https://www.databricks.com/trust/security-features

https://www.databricks.com/discover/data-governance

https://www.databricks.com/glossary/mlops

https://www.databricks.com/resources/whitepaper/databricks-ai-security-framework-dasf

Thank you

Ashish Pal