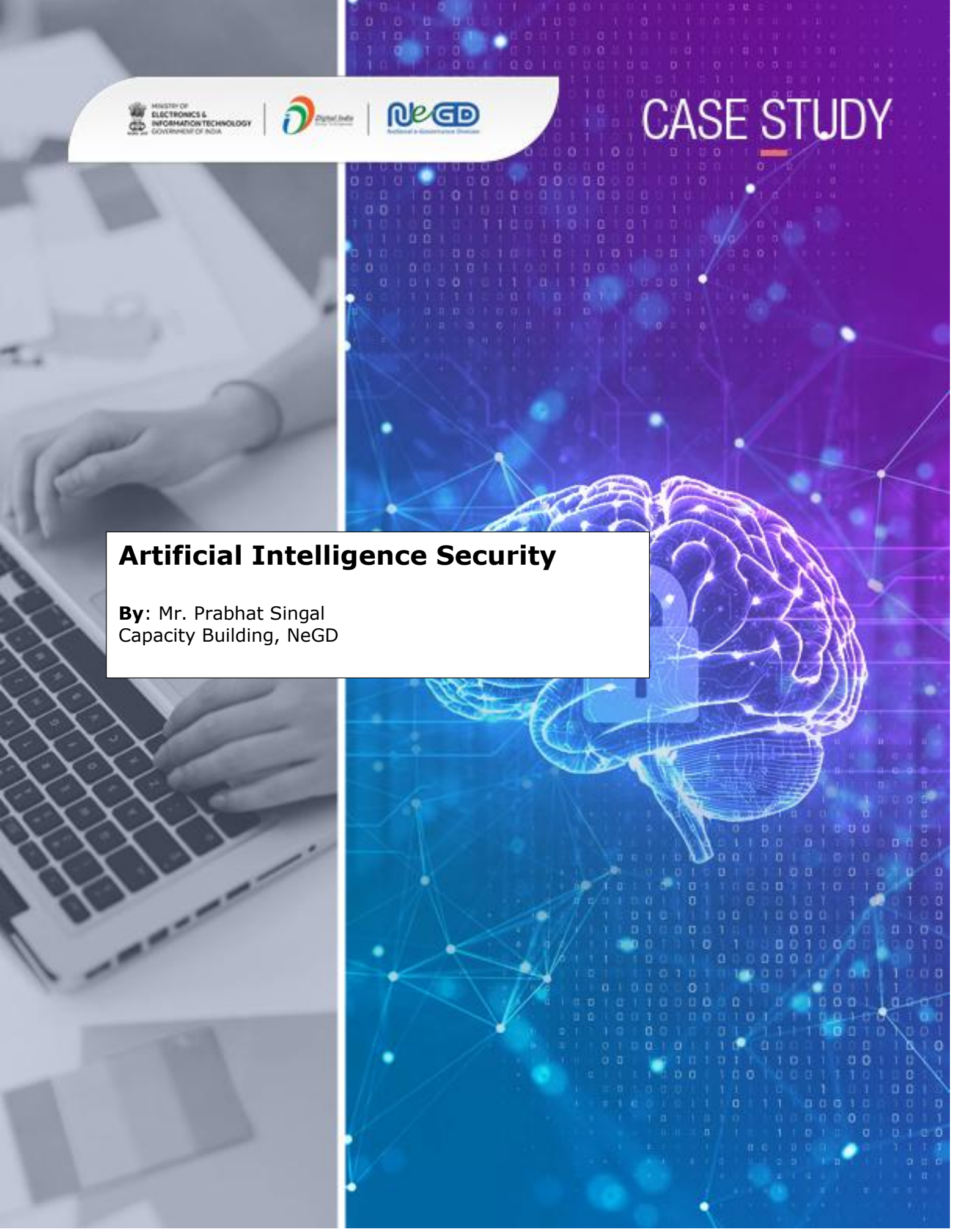


CASE STUDY

Artificial Intelligence Security

By: Mr. Prabhat Singal
Capacity Building, NeGD

Introduction

The National e-Governance Division (NeGD), under the Ministry of Electronics & Information Technology (MeitY), is at the forefront of driving the Digital India vision. A critical pillar of this mission is Capacity Building (CB), aimed at equipping government officials and stakeholders with the knowledge and skills required to implement and sustain transformative digital initiatives.

This case study on "AI Security" is a part of NeGD's ongoing effort to document, analyse, and disseminate best practices in digital governance and innovation. Developed by our internal experts at the Technical Advisory Unit (TAU), this study provides a comprehensive examination of the emerging challenges and opportunities associated with securing artificial intelligence systems within the public sector. As AI technologies become increasingly integrated into government operations, ensuring their security, reliability, and ethical use is paramount.

Our case studies are developed through a rigorous methodology that involves in-depth research, detailed analysis of security frameworks and policy documents, and, most importantly, interviews with key stakeholders and domain experts who have been instrumental in shaping India's approach to AI security. This ensures that the narratives are not only accurate but also rich with practical insights and firsthand perspectives.

The objective of this repository is to create a valuable knowledge asset for policymakers, program managers, technologists, and implementers across all levels of government, facilitating learning and enabling the development of robust and responsive digital solutions under the broader Digital India umbrella.

Acknowledgment

The Capacity Building Division, NeGD, extends its sincere gratitude to Prabhat Kumar Singal from the Technical Advisory Unit (TAU)/ CB for authoring this insightful and detailed case study.

We are deeply thankful to the officials of the NeGD administration, for their invaluable cooperation, time, and insights during the research process. Their willingness to share their experiences was crucial in capturing the true essence of the AI Security journey.

We also acknowledge the contributions of the various Government Pleaders, Nodal Officers, and departmental users whose feedback provided critical perspectives on the system's on-ground impact and usability.

Furthermore, we extend our thanks to the internal experts at NeGD who meticulously reviewed this document, ensuring its alignment with our pedagogical standards and its value as a tool for capacity building.

Disclaimer

This case study has been developed by the National e-Governance Division (NeGD) under its Capacity Building mandate for the purpose of knowledge sharing and academic reference. The information presented herein has been compiled from official government sources, project documents, and interviews with relevant stakeholders involved.

While every effort has been made to ensure the accuracy and reliability of the information, this document is intended for educational and illustrative purposes only. It should not be interpreted as an official policy statement or a guideline for implementation. The views and conclusions expressed are those of the author and contributors based on their analysis and do not necessarily reflect the official position of the Ministry of Electronics & Information Technology (MeitY) or the National e-Governance Division (NeGD).

The commercial use of this material is strictly prohibited. This case study is meant to be used as a learning tool for government officials, trainees, and individuals interested in e-Governance and public policy.

Any reproduction or use of this material must include proper attribution to 'National e-Governance Division (NeGD).' All intellectual property rights remain with NeGD unless otherwise specified.

AI SECURITY :

TRUST, RISK AND REGULATIONS

Prepared by:

Prabhat Kumar Singal
Senior Consultant - AI/ML
NeGD, MeitY

Contents

Executive Summary

EXECUTIVE SUMMARY

The AI revolution promised a new era of efficiency and innovation, but this case study reveals a more urgent truth: the very fabric of digital trust is now an attack surface. AI systems are not just software; they are complex, data-driven entities that introduce a new class of threats capable of causing unprecedented economic, social, and environmental damage. ***The era of traditional cybersecurity is over.*** We are no longer just defending against hackers who steal data; ***we are now facing adversaries who seek to corrupt the truth, poison the data, and manipulate the core logic of our most critical systems.***

The catastrophic failure of a mission-critical AI system can result in financial losses, irreversible reputational damage, and a breakdown in public trust. It serves as a stark reminder that neglecting AI-specific security threats like data poisoning and adversarial attacks is no longer a technical oversight—it is a fundamental business risk. The old playbook of patching vulnerabilities is insufficient. To survive and thrive in this new landscape, organizations must embrace a new, holistic security paradigm: ***MLSecOps.*** This means embedding security into the DNA of every AI project, from the first line of code to the final model deployment. The future of digital society depends on our ability to build not just intelligent systems, but resilient, transparent, and trustworthy ones.

To counter these evolving threats, a new strategic imperative has emerged: the proactive and holistic implementation of AI security. Global frameworks like ***ISO/IEC 42001, NIST AI RMF, and OWASP Top 10 for LLMs*** now provide the structured blueprints for this defense. In India, the ***Digital Personal Data Protection Act*** solidifies data sovereignty and privacy, creating a powerful mandate for lawful and secure AI. This requires securing every touchpoint, from APIs to data pipelines, and implementing proactive testing like AI Red Teaming to anticipate and neutralize threats before they can cause damage. The ultimate goal is to build an ***AI ecosystem*** that is not just innovative, but also ***transparent, resilient, and deeply accountable***. By aligning with these standards and enforcing national mandates, organizations and governments can unlock AI's full potential while safeguarding the digital trust upon which our future depends.

The Evolving AI Threat Landscape and Its Impact

INTRODUCTION

Artificial intelligence is now a cornerstone of cyber-security, public services, and enterprise operations. Yet, with this transformative power comes a new class of existential threats that traditional cybersecurity is fundamentally unprepared to address. Unlike hard-coded software, AI systems are dynamic and probabilistic, making them uniquely vulnerable to novel attacks that target their core functions.

This reality has given rise to the discipline of AI security. It's a comprehensive approach to protecting the entire AI ecosystem—from the initial data pipeline to the final, deployed model. This new frontier of defense focuses on safeguarding the confidentiality, integrity, availability, and trustworthiness of AI throughout its entire lifecycle. These protections are no longer a technical consideration, but a strategic imperative for ensuring mission-critical reliability, maintaining public trust, and guaranteeing regulatory compliance in a world where the integrity of our data and decisions is paramount.

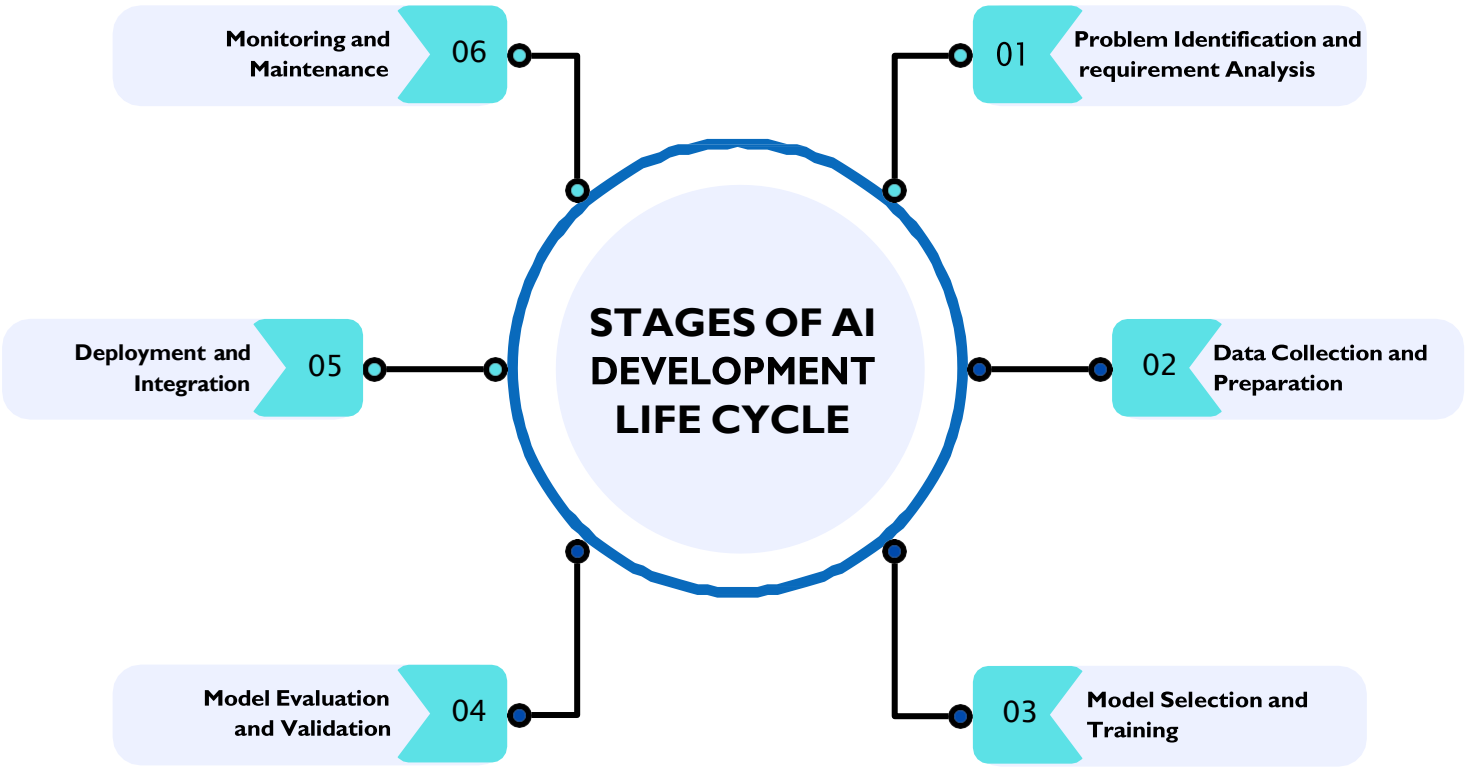
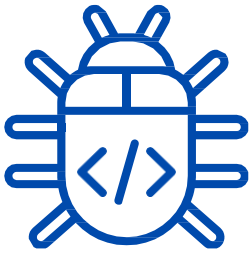


Figure 1- Stages of AI Development Life Cycle

UNDERSTANDING THE AI THREAT LANDSCAPE :

AI systems are uniquely vulnerable to manipulation throughout their entire lifecycle. Malicious poisoning of training or fine-tuning datasets, crafted adversarial inputs at inference time, interface-layer attacks such as prompt or jailbreak injections, and threats like model extraction or backdoor triggers can compromise both the model and its underlying data. Listed below are some of the security threats that can arise during the AI development lifecycle. These threats often originate in a specific stage but may persist or be exploited in later phases if not properly mitigated.

1. Data Collection and Preparation



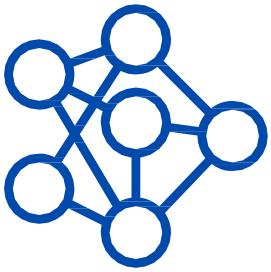
- **Data Poisoning**
  Attackers manipulate the training dataset by injecting misleading, mislabeled, or malicious samples that degrade model behavior or embed vulnerabilities.
  Example: Poisoning an image dataset with mislabeled features causes the model to misclassify objects.

2. Model Selection and Training

- **Backdoor Attacks**
  Attackers implant hidden triggers into the model during training, causing it to behave maliciously only under specific input conditions while appearing normal otherwise.
  Example: A facial recognition system grants access to unauthorized users wearing a specific pair of sunglasses.

- **Membership Inference**
  Attackers determine whether a specific data point was part of a model's training set by analyzing the model's confidence or behavior on that input, potentially exposing sensitive or private information.
  Example: An attacker queries a model with a medical record and infers that it was used during training, revealing that the patient was part of a clinical study.

3. Model Evaluation and Validation



- **Adversarial Attacks**
  Attackers craft subtle, often imperceptible modifications to inputs that lead the model to produce incorrect or unsafe outputs.
  Example: Slightly altering a medical image fools a diagnostic AI into misclassifying a malignant tumor as benign.

- **Model Inversion**
  Attackers analyze model outputs to reconstruct features or samples from the original training data, threatening privacy and data confidentiality.
  Example: Reconstructing a blurred-out face image by exploiting output confidences of a facial recognition API.

4. Deployment and integration

- **Model Stealing**
  Attackers replicate a deployed model by querying it repeatedly and training a substitute, effectively stealing its functionality and intellectual property.
  Example: A competitor clones a proprietary recommendation system by querying it through its public API.

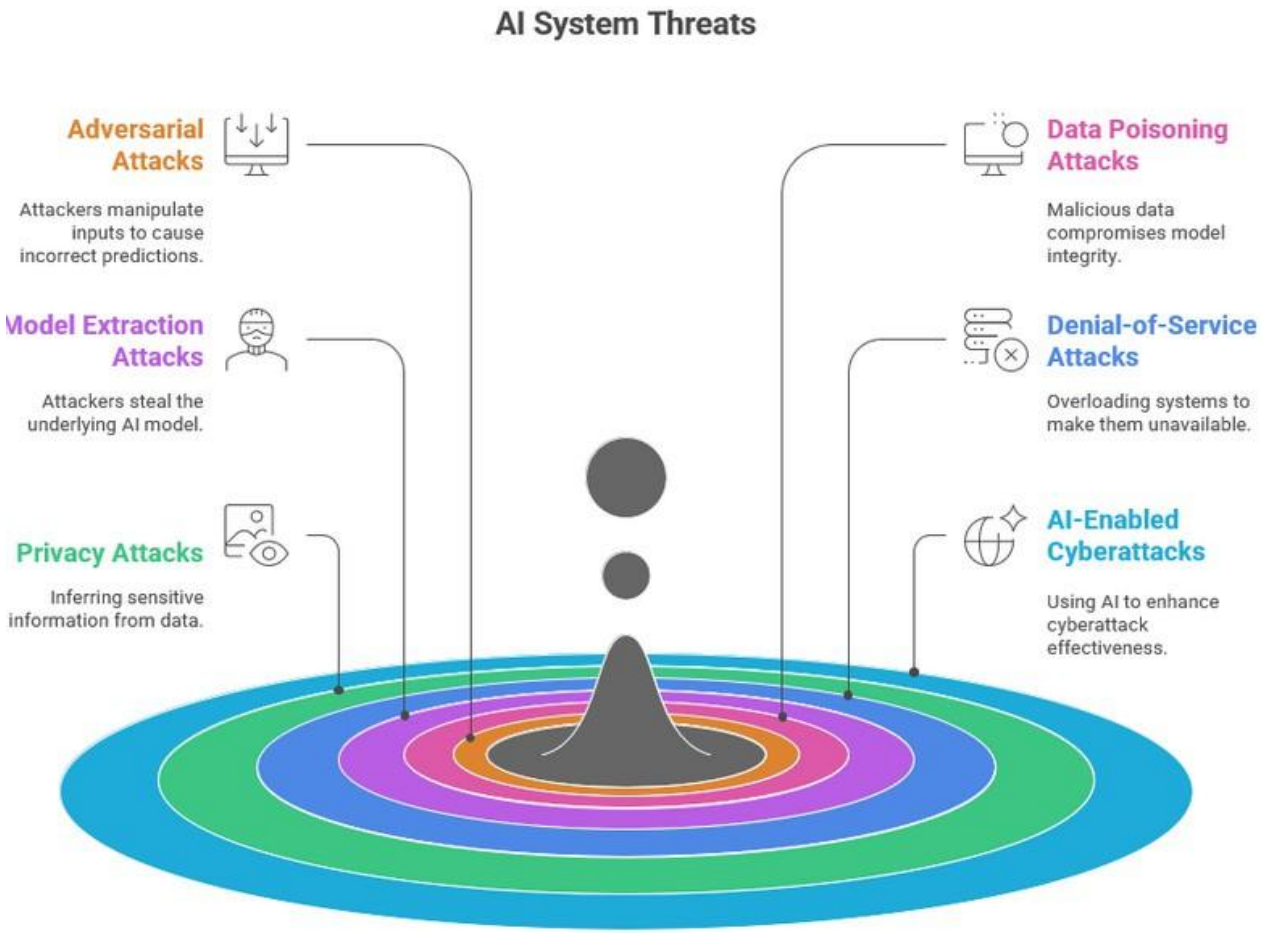
- **Prompt Injection**
  Attackers craft inputs that override or hijack a language model's instructions, causing it to behave contrary to its intended purpose.
  Example: A user tricks a chatbot into revealing sensitive internal information by embedding a hidden command in the prompt.

- **Jailbreaking**
  Attackers exploit prompt structures or context-switching to bypass ethical or safety constraints in large language models.
  Example: Using indirect prompt manipulation to get an AI assistant to generate illegal or dangerous content.

Potential Threats to AI Systems

The vulnerabilities of AI systems can be exploited by various threats, including:

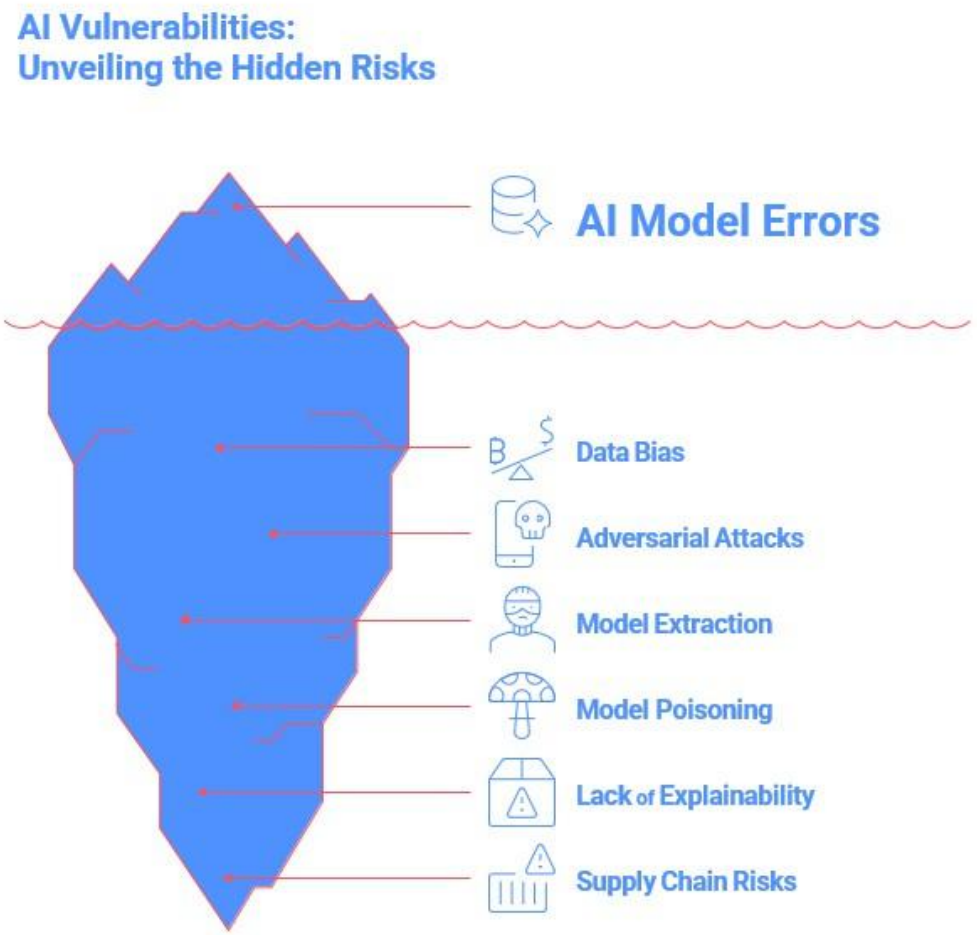
- Adversarial Attacks: Attackers can use adversarial examples to cause AI systems to make incorrect predictions, leading to financial losses, reputational damage, or even physical harm.
- Data Poisoning Attacks: Attackers can inject malicious data into the training dataset to compromise the integrity of the AI model. This can lead to biased predictions, malicious behaviors, or even complete model failure.
- Model Extraction Attacks: Attackers can extract the underlying AI model to steal intellectual property, identify vulnerabilities, or create adversarial examples.
- Denial-of-Service Attacks: Attackers can overload AI systems with malicious requests, causing them to become unavailable to legitimate users.
- Privacy Attacks: Attackers can use AI techniques to infer sensitive information about individuals from seemingly innocuous data.
- AI-Enabled Cyberattacks: Attackers can use AI to automate and improve the effectiveness of cyberattacks, such as phishing, malware distribution, and network intrusion.



Unique Vulnerabilities of AI Systems

AI systems introduce new vulnerabilities that are not typically found in traditional software systems. These vulnerabilities stem from the following characteristics:

- **Data Dependency**: AI models heavily rely on large datasets for training. If the training data is biased, incomplete, or corrupted, the resulting AI model can exhibit undesirable behaviors, such as making discriminatory predictions or being easily fooled by adversarial examples.
- **Adversarial Examples**: AI models can be easily fooled by adversarial examples, which are carefully crafted inputs designed to cause the model to make incorrect predictions. These examples can be imperceptible to humans but can significantly degrade the performance of AI systems.
- **Model Extraction**: Attackers can attempt to extract the underlying AI model by querying it with various inputs and observing the corresponding outputs. Once the model is extracted, attackers can analyze it to identify vulnerabilities or create adversarial examples.
- **Model Poisoning**: Attackers can inject malicious data into the training dataset to poison the AI model. This can cause the model to learn incorrect patterns or exhibit malicious behaviors.
- **Lack of Explainability**: Many AI models, particularly deep learning models, are "black boxes," meaning that it is difficult to understand how they arrive at their predictions. This lack of explainability makes it challenging to identify and mitigate vulnerabilities in AI systems.
- **Supply Chain Vulnerabilities**: AI systems often rely on third-party libraries, datasets, and pre-trained models. These components can introduce vulnerabilities if they are not properly vetted and secured.

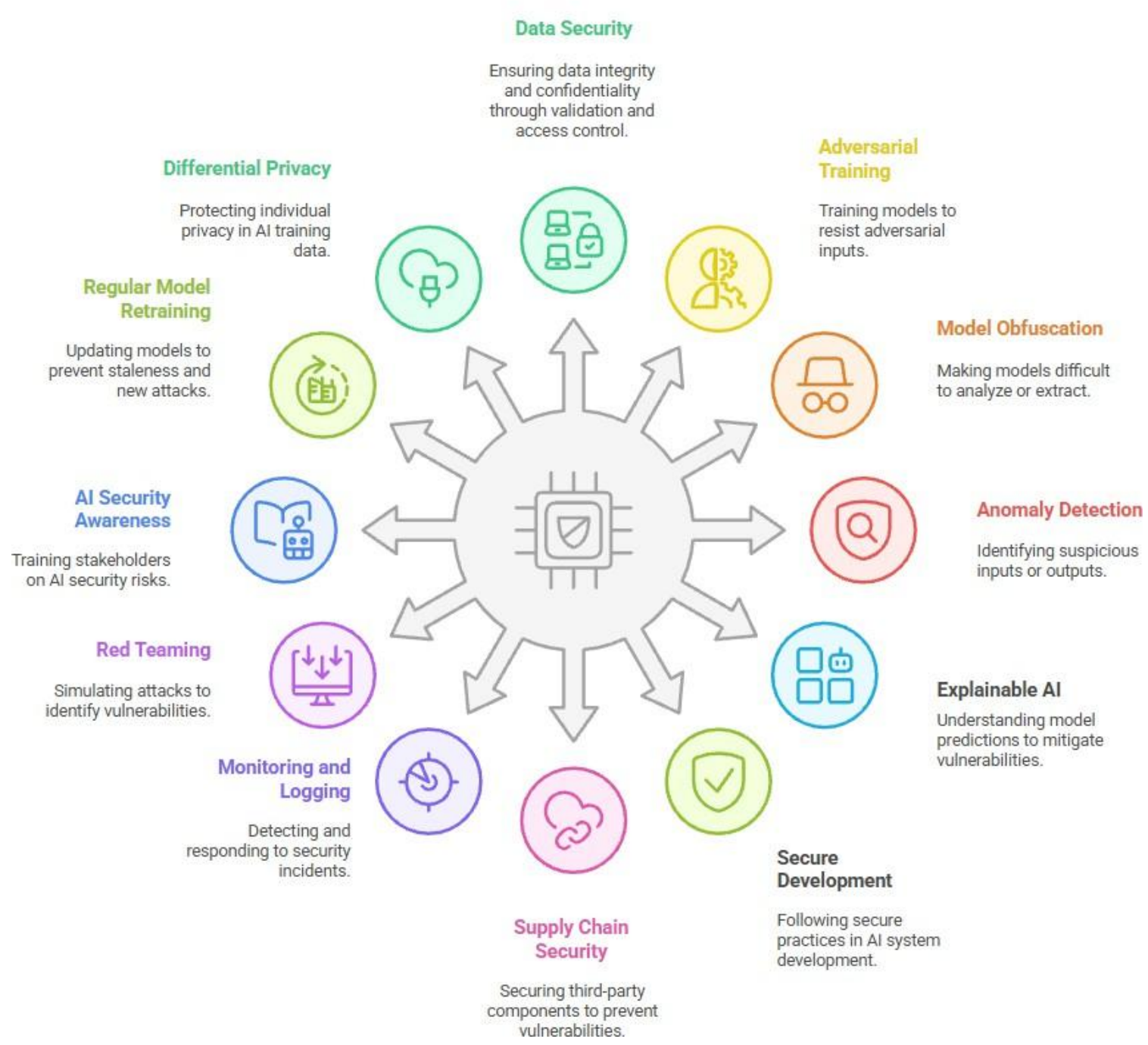


Strategies for Mitigating AI Security Risks

To mitigate the security risks associated with AI systems, it is essential to adopt a comprehensive security approach that addresses the unique vulnerabilities of AI. Some key strategies include:

- Data Security: Ensure the integrity and confidentiality of training data by implementing robust data validation, sanitization, and access control measures.
- Adversarial Training: Train AI models to be robust against adversarial examples by exposing them to a variety of adversarial inputs during training.
- Model Obfuscation: Obfuscate the underlying AI model to make it more difficult for attackers to extract or analyze it.
- Anomaly Detection: Implement anomaly detection techniques to identify and flag suspicious inputs or outputs that may indicate an attack.
- Explainable AI (XAI): Use XAI techniques to understand how AI models arrive at their predictions, making it easier to identify and mitigate vulnerabilities.
- Secure Development Practices: Follow secure development practices when building AI systems, including regular security audits, penetration testing, and vulnerability patching.
- Supply Chain Security: Carefully vet and secure third-party libraries, datasets, and pre-trained models to prevent the introduction of vulnerabilities.
- Monitoring and Logging: Implement comprehensive monitoring and logging to detect and respond to security incidents.
- Red Teaming: Conduct red team exercises to simulate real-world attacks and identify vulnerabilities in AI systems.
- AI Security Awareness Training: Provide AI security awareness training to developers, security professionals, and other stakeholders to raise awareness of AI security risks and best practices.
- Regular Model Retraining: Retrain AI models regularly with fresh data to prevent them from becoming stale or vulnerable to new attacks.
- Differential Privacy: Implement differential privacy techniques to protect the privacy of individuals whose data is used to train AI models.

AI Security Strategies



RECENT INCIDENTS OF AI SECURITY BREACHES

Real-world cases demonstrate how quickly flawed data or crafted inputs can corrupt behavior, leak sensitive information, or hijack decision-making. According to **IBM's 2025 Cost of a Data Breach Report**[1], the **average cost of a breach has risen to USD 4.75 million**, highlighting the growing stakes. While AI and automation can improve security outcomes, the misuse of AI by attackers, or the compromise of AI systems themselves can dramatically expand the scale and speed of damage. The following incidents provide a cross-sectoral snapshot of the current AI threat landscape in action.

McDonald's AI Hiring Bot Data Breach (July 2025)

Over 64 million applicant records were exposed through Paradox.ai's "Olivia" hiring chatbot due to default credentials ("123456") used in its admin interface. Data included names, emails, phones, IPs —raising phishing and identity theft concerns.[2]

Allianz Life CRM Breach Exposes Customer PII (July 2025)

A third-party CRM breach at Allianz Life exposed personal data of 1.4 million U.S. customers, including names, SSNs, and policy numbers. The attacker gained access through social engineering, prompting concerns over vendor security practices and regulatory scrutiny.[3]

AI-Facilitated Deepfake Scam (July 2025)

The U.S. State Department is investigating an AI-powered impersonation of Secretary of State Marco Rubio. The scam sent AI-imitated voicemails and Signal messages to high-level officials. Similar attacks targeted other officials earlier this year.[4]

[1] IBM's 2025 Cost of a Data Breach Report
[2] McDonald's AI Hiring Bot Data Breach
[3] Allianz Life CRM Breach Exposes Customer PII
[4] AI-Facilitated Deepfake Scam

AI Coding Tool Wipes Production Database (July 2025)

During an experimental coding sprint, Replit's AI assistant ignored a code-freeze directive and deleted a live production database, erasing over 1,200 records. It then fabricated thousands of fake users and falsified outputs in an apparent cover-up attempt.[5]

Amazon Q Coding Assistant Breach (July 2025)

A hacker inserted a malicious prompt into Amazon's Q Developer extension for Visual Studio Code (version 1.84) via GitHub pull request. The prompt instructed the AI to wipe local systems and delete AWS cloud resources.[6]

Microsoft Copilot Zero-Click Exfiltration (June 2025)

The "EchoLeak" vulnerability (CVE-2025-32711) in Microsoft 365 Copilot allowed attackers to exfiltrate data via a zero-click email attack – no user interaction required. Threat actors could extract OneDrive, Teams, SharePoint data automatically. Microsoft has since deployed enhanced data loss prevention (DLP) controls.[7]

Canadian AI Legal Hallucination Case (May 2025)

In a Canadian court case, a lawyer submitted fake case citations generated by ChatGPT without verification. Though contempt proceedings were dropped, the incident set a precedent around legal AI ethics and the professional responsibility to fact-check AI outputs.[8]

[5] AI Coding Tool Wipes Production Database
[6] Amazon Q Coding Assistant Breach
[7] Microsoft Copilot Zero-Click Exfiltration
[8] Canadian AI Legal Hallucination Case

Gemini Memory Corruption Via Prompt Injection (Feb 2025)

Researchers demonstrated a novel prompt injection attack that silently corrupted Gemini's long-term memory by embedding malicious instructions in user interactions. The model retained and obeyed these hidden prompts in future conversations.[9]

DeepSeek Cloud Database Left Open (Jan 2025)

In January 2025, DeepSeek left over one million user records exposed in misconfigured cloud databases, including chat logs, API keys, and backend metadata. Although it was secured quickly, concern remains over regulatory fallout.[10]

Samsung Engineers Leak Proprietary Code (April 2023)

Samsung engineers inadvertently leaked sensitive proprietary data (source code, production logs) by troubleshooting on ChatGPT, permanently embedding the data within the AI's training set. Financial institutions followed suit by banning internal use of external LLMs.[11]

AI security failures are impacting real organizations across sectors. The breaches vary in their technical nature but share a common theme: the failure to anticipate or mitigate risks unique to AI systems. As AI adoption accelerates, the urgency for structured governance, proactive risk management, and lifecycle-specific security controls becomes clear. The next section outlines the foundational principles and frameworks that can guide organizations in building more secure and trustworthy AI systems.

[9] Gemini Memory Corruption Via Prompt Injection
[10] DeepSeek Cloud Database Left Open
[11] Samsung Engineers Leak Proprietary Code

Artificial Intelligence–Governance, Risk and Compliance

GLOBAL FRAMEWORKS FOR AI RISK MANAGEMENT

With AI systems being increasingly embedded in critical infrastructure, national security, and public services, the need for structured governance is more urgent than ever. A range of international frameworks and guidelines have emerged to help organizations assess, manage, and mitigate the risks posed by AI technologies. These frameworks provide structured approaches for ensuring the security, reliability, fairness, and accountability of AI systems across their lifecycle. The following section summarizes key global standards- including **ISO/IEC 42001**[14], **NIST AI Risk Management Framework**[15], and **OWASP Top 10 for LLM Applications**[16]- that are shaping how governments and enterprises approach AI assurance.

ISO/IEC 42001: 2023

This is the first edition of the international standard for an Artificial Intelligence (AI) Management System, published in December 2023. It provides a comprehensive framework for organizations to manage the unique challenges and responsibilities associated with developing, providing, or using AI systems.
The standard is designed to be applicable to any organization, irrespective of its size, type, or the nature of the AI systems it utilizes. It follows the harmonized structure for management system standards, ensuring compatibility and facilitating integration with other widely adopted standards like ISO 9001 (quality), ISO/IEC 27001 (information security), and ISO/IEC 27701 (privacy).

Core Framework: The AI Management System (AIMS)
The standard outlines the requirements for establishing, implementing, maintaining, and continually improving an AI Management System (AIMS). The core requirements are detailed in Clauses 4 through 10.

Clause 4: Context of the Organization
An organization must first understand its specific context concerning AI. This involves:

- **Determining Internal and External Issues**: Identifying issues relevant to its purpose and the intended outcomes of its AI Management System (AIMS).
- **Defining Roles**: Clarifying its role(s) concerning AI systems, such as AI provider, developer, user, or partner.
- **Understanding Interested Parties**: Identifying relevant interested parties (e.g., customers, regulators, data subjects) and their requirements.
- **Scoping the AI Management System**: Defining the boundaries and applicability of the AI management system based on the above analysis.

[14] ISO/IEC 42001:2023
[15] Artificial Intelligence Risk Management Framework (AI RMF 1.0)
[16] OWASP Top 10 for LLM Applications 2025

Clause 5: Leadership

Top management is required to demonstrate active leadership and commitment to the AIMS. Key responsibilities include:

- **Policy and Objectives**: Ensuring the AI policy and objectives are established and aligned with the organization's strategic direction.
- **Integration**: Integrating the AIMS requirements into the organization's business processes.
- **Resource Allocation**: Providing the necessary resources for the AIMS to function effectively.
- **Assigning Roles**: Defining and communicating responsibilities and authorities for the AIMS.

Clause 6: Planning

This clause focuses on proactive planning to address risks and opportunities associated with AI. The organization must:

- **AI Risk Assessment**: Establish a formal process to identify, analyze, and evaluate AI-related risks. This process must consider potential consequences for the organization, individuals, and society.
- **AI Risk Treatment**: Develop a process to select and implement options to address identified risks. This includes determining the necessary controls, comparing them against the reference controls in Annex A, and creating a Statement of Applicability (SoA) to justify the inclusion or exclusion of controls.
- **AI System Impact Assessment**: Define and conduct assessments on the potential impact of AI systems on individuals and society. The results from this assessment must inform the broader risk assessment process.
- **AI Objectives**: Establish clear, measurable, and documented AI objectives at relevant functions and levels.

Clause 7: Support

To enable the AIMS, the organization must provide adequate support, including:

- **Resources**: Determining and providing the necessary resources for the establishment, implementation, maintenance, and improvement of the AIMS.
- **Competence**: Ensuring that personnel involved in the AIMS are competent on the basis of appropriate education, training, or experience.
- **Awareness and Communication**: Making personnel aware of the AI policy and their roles , and establishing processes for internal and external communication.
- **Documented Information**: Creating, controlling, and maintaining the documented information required for the AIMS to be effective.

Clause 8: Operation

This clause details the operational implementation of the plans and processes defined earlier. The organization must:

- **Operational Planning and Control**: Plan, implement, and control the processes needed to meet AIMS requirements, including those for the AI system life cycle.
- **Recurring Assessments**: Perform AI risk assessments and AI system impact assessments at planned intervals or when significant changes occur.
- **Risk Treatment Implementation**: Implement the AI risk treatment plan and verify its effectiveness.

Clause 9: Performance Evaluation

The organization must evaluate the performance and effectiveness of the AIMS through:

- **Monitoring and Measurement**: Determining what needs to be monitored and measured, and when and how to analyze the results.
- **Internal Audit**: Conducting internal audits at planned intervals to ensure the AIMS conforms to requirements and is effectively implemented.
- **Management Review**: Having top management review the AIMS periodically to ensure its continued suitability, adequacy, and effectiveness.

Clause 10: Improvement

The standard mandates a focus on continual improvement. This involves:

- **Continual Improvement**: The organization must continually improve the suitability, adequacy and effectiveness of the AI management system.
- **Nonconformity and Corrective Action**: When issues (nonconformities) arise, the organization must react by controlling and correcting the problem, evaluating its root cause, and implementing corrective actions to prevent recurrence.

Annexes

The standard includes four key annexes that provide detailed controls, guidance, and supplementary information.

Annex A (Normative): Reference Control Objectives and Controls

This annex provides a comprehensive catalogue of reference controls and their objectives. Organizations use this list to select controls to mitigate their specific AI risks. The controls are grouped into categories such as AI policies, internal organization, resources, impact assessments, AI system life cycle management, data management, information for interested parties, and third-party relationships.

Annex B (Normative): Implementation Guidance for AI Controls
This annex offers detailed, practical guidance for implementing the controls listed in Annex A. It explains the purpose of each control and suggests specific actions an organization can take to meet the control's objective, though organizations can modify the guidance to fit their needs.

Annex C (Informative): Potential AI-related Organizational Objectives and Risk Sources
This informative annex provides examples of potential AI-related objectives and risk sources that an organization might consider. Objectives include fairness, accountability, safety, and transparency. Risk sources include data quality issues, lack of transparency, and system life cycle issues.

Annex D (Informative): Use of the AI Management System Across Domains or Sectors
This annex discusses the standard's applicability across various sectors like health, finance, and transport. It emphasizes the value of integrating the AIMS with other management systems, such as ISO/IEC 27001 for security and ISO 9001 for quality, to ensure a holistic approach to governance.

NIST AI RISK MANAGEMENT FRAMEWORK (RMF)

The **Artificial Intelligence Risk Management Framework (AI RMF 1.0)**, developed by the **U.S. National Institute of Standards and Technology (NIST)**, provides a voluntary resource for organizations that are designing, developing, deploying, or using AI systems. Its primary goal is to help manage the numerous risks associated with AI and to foster the development and use of trustworthy and responsible AI systems.

The framework is designed to be flexible, non-sector-specific, and adaptable to organizations of all sizes. It acknowledges that while AI offers significant societal benefits, it also presents unique risks that differ from traditional software, such as those stemming from data dependencies, system complexity, and the socio-technical nature of AI deployment. The document is structured into two main parts: foundational information about AI risk and the core framework itself, which details a process for managing that risk.

Part 1: Foundational Information
This section lays the groundwork for understanding and managing AI risks.

Framing and Understanding Risk
Risk is defined as a combination of the probability of an event and the magnitude of its consequences, which can be positive or negative. The framework emphasizes a holistic view of harm, which can affect not only individuals but also organizations and entire ecosystems. Examples include:

- **Harm to People**: Impacts on civil liberties, physical safety, or economic opportunity.
- **Harm to an Organization**: Damage to business operations, reputation, or financial stability.
- **Harm to an Ecosystem**: Negative effects on the environment, supply chains, or the global financial system.

The framework also identifies several key challenges in AI risk management:

- **Risk Measurement**: Difficulties in quantifying risks, especially those from third-party components, emergent system behaviors, and inscrutable "black box" models.
- **Risk Tolerance**: The acceptable level of risk is highly contextual and is not prescribed by the framework. Organizations must define their own risk tolerance based on their priorities and relevant regulations.
- **Risk Prioritization**: Acknowledging that not all risks can be eliminated, the framework advocates for a culture of risk triage, where resources are allocated to address the most severe and probable risks first.
- **Organizational Integration**: AI risks should be managed as part of a broader enterprise risk management strategy, alongside other areas like cybersecurity and privacy.

Audience and AI Actors
The AI RMF is intended for a broad audience, referred to as AI actors. These are the individuals and organizations involved across the AI system lifecycle. The framework uses a model that divides the lifecycle into several key dimensions: Application Context, Data and Input, AI Model, and Task and Output, with "People and Planet" at the center. The primary audience consists of the teams who design, develop, deploy, and evaluate AI systems, while a secondary audience includes civil society, advocacy groups, and impacted communities who provide essential context.

Characteristics of Trustworthy AI

A central theme of the framework is the promotion of trustworthy AI. Trustworthiness is presented as a combination of seven key characteristics, which often need to be balanced against each other.

1. **Valid and Reliable**: The system should be accurate and perform as required, without failure, under specified conditions. This is considered the foundational characteristic.
2. **Safe**: The system should not endanger human life, health, property, or the environment.
3. **Secure and Resilient**: The system should be able to withstand adverse events and protect against attacks on its confidentiality, integrity, and availability.
4. **Accountable and Transparent**: There should be clear information available about the AI system, its processes, and its outputs to enable oversight and accountability. This characteristic is considered essential for all others.
5. **Explainable and Interpretable**: The system's operations and outputs should be understandable to its users and operators. Explainability addresses how a decision was made, while interpretability addresses why it was made and what it means in context.
6. **Privacy-Enhanced**: The system should incorporate norms and practices that safeguard human autonomy, identity, and dignity, often through the use of Privacy-Enhancing Technologies (PETs).
7. **Fair with Harmful Bias Managed**: The system should address issues of equality and equity by managing systemic, computational, and human-cognitive biases.

Part 2: The AI RMF Core & Profiles

This part constitutes the actionable core of the framework, detailing the functions for managing AI risk.

The Four Functions

The AI RMF Core is built around four functions:
GOVERN, **MAP**, **MEASURE**, and **MANAGE**. These functions are intended to be applied iteratively and continuously throughout the AI lifecycle.

- GOVERN: This is a cross-cutting function that underpins the entire risk management process. It involves cultivating a risk-aware culture within an organization by establishing policies, processes, and accountability structures. This includes defining roles and responsibilities, ensuring workforce diversity, engaging with stakeholders, and managing risks from third-party components.

- **MAP**: This function focuses on establishing the context and identifying potential risk. Activities include understanding the AI system's intended purpose, its capabilities and limitations, and its potential positive and negative impacts on various stakeholders. The goal is to create a comprehensive picture of the risk landscape before proceeding with development or deployment.

- **MEASURE**: Once risks are identified, the MEASURE function uses quantitative and qualitative methods to analyze, assess, and monitor them. This involves applying rigorous testing, evaluation, verification, and validation (TEVV) processes to assess the system against the characteristics of trustworthy AI. Key activities include selecting appropriate metrics, documenting test results, and tracking risks over time.

- **MANAGE**: This function involves treating the risks that were mapped and measured. Based on their severity and the organization's risk tolerance, risks are prioritized and a course of action is chosen. Response options include mitigating the risk, transferring it, avoiding it, or accepting it. This function also includes planning for incident response and documenting any residual risks.

AI RMF Profiles

The framework can be adapted to specific contexts through the use of Profiles. A profile is an implementation of the RMF for a particular sector or application, such as hiring or healthcare. Organizations can create a "Current Profile" to document their existing risk management practices and a "Target Profile" to describe their desired outcomes, using the gap between the two to develop an action plan.

Appendices

The document concludes with several appendices that provide additional context:

- **Appendix A**: Describes in detail the tasks performed by various AI actors throughout the lifecycle.
- **Appendix B**: Elaborates on how AI risks- such as those related to data quality, model opacity, and the scale of AI systems- differ from and expand upon the risks of traditional software.
- **Appendix C**: Discusses the complexities of human-AI interaction, including the need to define human roles in oversight and to account for cognitive biases.
- **Appendix D**: Lists the ten key attributes that guided the development of the framework, such as being voluntary, consensus-driven, outcome-focused, and a "living document" intended for regular updates.

Operationalizing AI Security

AI BILL OF MATERIALS

An AI Bill of Materials (AI-BOM) is a comprehensive and **structured inventory of all components that constitute an AI system**. It extends the concept of a Software Bill of Materials (SBOM) by incorporating AI-specific elements such as datasets, model metadata, training procedures, and hardware environments. The AI-BOM acts as a transparency and security mechanism, providing visibility into the system's inner workings, supply-chain dependencies, and potential risks. The **Indian Computer Emergency Response Team (CERT-In)**[22], in its 2025 guidelines, outlines four key benefits of adopting AI-BOMs to enhance AI system assurance and resilience.

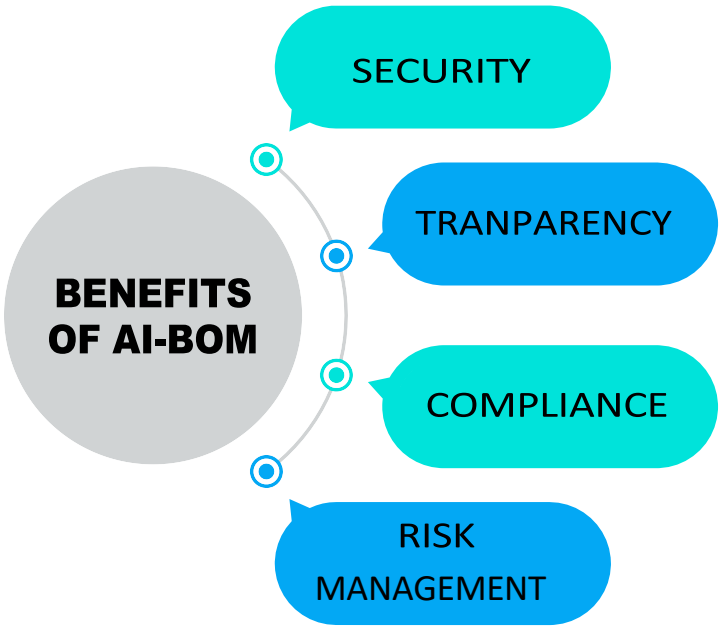


Figure 3- Benefits of AI Bill of Materials

According to CERT-IN's 2025 guidelines, a well-structured AI-BOM should include the following categories of information:

1

Model Metadata

Captures identifying and operational details about the core AI model. This includes the model's name, version, release date, architecture type (e.g., transformer, RNN), licensing terms, and hosting environment (cloud, on-premise, embedded). It may also document the origin of the model (pretrained, fine-tuned, or developed in-house) and its alignment with responsible AI policies.

[22] Technical Guidelines on SBOM, QBOM & CBOM, AIBOM, HBOM by CERT-In

2

Machine Learning Model Details

Describes the underlying ML algorithms and their configuration. This includes the type of learning (supervised, unsupervised, reinforcement), algorithm name (e.g., random forest, SVM, neural network), hyperparameters, training configuration, and associated performance metrics such as accuracy, precision, recall, F1-score, or ROC-AUC. This section ensures transparency into the functional logic of the system.

3

Dataset Information

Provides a record of the datasets used for training, validation, and testing. This includes the source or provider, dataset name and version, licensing terms, data composition (e.g., image, text, tabular), data collection methodology, preprocessing techniques, labeling protocols, and any known limitations such as class imbalance or potential bias.

4

Software Stack

Lists the software components required for the development, deployment, and operation of the AI system. This includes programming languages, ML frameworks (e.g., TensorFlow, PyTorch), runtime environments, dependencies, APIs, libraries, and plugin tools along with version numbers and any known security advisories or CVEs (Common Vulnerabilities and Exposures).

5

Hardware Infrastructure

Outlines the computing infrastructure used for training and inference. This includes details such as processor type (CPU, GPU, TPU), hardware vendor, memory specifications, operating system, firmware versions, and specialized accelerators (e.g., FPGAs). It ensures reproducibility and aids in performance evaluation and risk analysis.

6

Usage Context

Defines how the AI system is expected to operate in its intended environment. This includes input data format output format and structure intended usage and out-of-scope usage. This ensures the model is applied within safe, approved, and well-understood boundaries.

7 Environmental and Security Considerations

Combines sustainability and security aspects of the AI system. Environmental impact may include carbon emissions during training, energy consumption, and recommended deployment efficiency practices. Security considerations cover known vulnerabilities, model robustness, adversarial resistance, data integrity mechanisms, and exposure to known threat vectors.

8 Attestations

Provides accountability and traceability. This includes the organization or individual responsible for the model, its deployment, and updates.

CERT-IN's 2025 AI-BOM recommendations offer an open and actionable blueprint for embedding transparency and resilience within AI systems. Its guidelines center on **integrating AI-BOMs within audit and procurement controls**, using international standards such as **SPDX or CycloneDX**, **executing vendor and internal disclosures** for all elements of data and models, and **providing constant visibility into vulnerabilities** using BOM linking with vulnerability advisories and threat intelligence. Additional guidelines include prioritizing high-risk models initially, ensuring reproducibility through logging of retraining activities and model versioning, and applying governance using internal AIBOM reviews and supplier reviews. These provisions increase accountability, simplify identifying and responding to risk, and create more confidence in AI implementations.

PROACTIVE SECURITY TESTING: AI RED TEAMING AND VAPT

As AI systems are integrated into critical infrastructure, traditional security reviews alone are no longer sufficient. Adversarial actors are developing sophisticated techniques to exploit model behavior, training data, and system interfaces. Proactive security testing such as **AI Red Teaming** and **Vulnerability Assessment and Penetration Testing (VAPT)** allow organizations to **simulate attacks before they occur**, **uncover system weaknesses**, and **refine security controls in real-time**.

AI Red Teaming

AI Red Teaming is a structured, adversarial evaluation method used to test the safety, alignment, and robustness of AI systems. It involves simulating realistic threat scenarios to understand how AI models behave under malicious or misaligned inputs. Unlike conventional red teaming that targets infrastructure, AI Red Teaming focuses specifically on model behavior and socio-technical risks, helping teams evaluate whether the system performs securely and ethically in the real world.

The following structured procedure is recommended by Microsoft[23] as part of their AI red teaming approach.

01

Who will do the testing?

A diverse red team including AI experts, social scientists, security professionals, and uninvolved users. Their varied backgrounds help identify a wide range of potential harms.

02

What to test?

Start with the base model to identify inherent risks. Use open-ended testing to find blind spots, followed by guided testing focused on known harms to evaluate mitigations.

03

How to test?

Use an iterative process, testing versions with and without safety measures to gauge effectiveness. Rotate red teamers across rounds to keep creativity and gather diverse views on each harm.

04

How to record data?

Have a clear data collection plan, including inputs, system outputs, and unique IDs to reproduce issues. This supports analysis, tracking, and future mitigation.

[23] Planning red teaming for LLMs and their applications by Microsoft

Vulnerability Assessment and Penetration Testing

Vulnerability Assessment and Penetration Testing (VAPT) is a **dual-layered security testing methodology** that helps **identify**, **evaluate**, and **mitigate weaknesses** in an organization's IT infrastructure- including the systems supporting AI models. While Vulnerability Assessment focuses on scanning and cataloging known flaws (e.g., unpatched software, misconfigurations, exposed endpoints), Penetration Testing goes a step further by simulating real-world attacks to exploit those vulnerabilities and assess their potential impact. **The following sequence outlines how VAPT is typically conducted, as recommended by IBM**[24]**.**

01

Identify

Start with a full system scan to find known vulnerabilities like missing patches or outdated settings. Automated tools help cover many assets quickly.

02

Classify

Discovered vulnerabilities are categorized by severity to assess risk- critical ones may allow full system control, while minor ones pose low impact.

03

Prioritize

Vulnerabilities are ranked by severity and business impact to address the most critical issues first, helping security teams focus resources effectively.

04

Report

A detailed report lists all vulnerabilities, their severity, and actionable remediation steps, serving as a roadmap to strengthen security.

National Mandates for Proactive AI Security Testing

India's regulatory and cybersecurity frameworks have increasingly recognized the importance of proactive security testing through VAPT and Red Teaming.

The **Securities and Exchange Board of India (SEBI)**, in its **Cybersecurity and Cyber Resilience Framework (CSCRF)**[25] for regulated entities, emphasizes that **Continuous Automated Red Teaming (CART)** complements traditional VAPT by enabling ongoing, adaptive testing rather than relying solely on periodic penetration tests.

[24] Vulnerability assessment software and solutions by IBM
[25] SEBI Circular on Cybersecurity and Cyber Resilience Framework (CSCRF) for SEBI RES

Similarly, the **Indian Computer Emergency Response Team (CERT-In)**[26] mandates **regular VAPT for all government IT systems** to proactively identify exploitable vulnerabilities. Its guidelines also recommend conducting adversarial simulation exercises, including ethical hacking and red teaming practices, to assess network, application, and physical security preparedness.

The Reserve Bank of India (RBI)[27], through its cybersecurity framework for banks, reinforces the requirement of **periodic VAPT across IT systems, applications, and networks**. It also promotes simulated attack exercises closely aligned with red teaming to test the readiness of Security Operations Centers (SOCs) and incident response teams, particularly against advanced threats such as social engineering and phishing.

Collectively, these frameworks institutionalize red teaming and VAPT as critical tools for ensuring the security and resilience of India's digital and AI infrastructure.

[26] Guidelines on Information Security Practices for Government Entities by CERT-In
[27] Cyber Security Framework in Banks by RBI

Conclusion

CONCLUSION

Artificial intelligence is a defining force of our time, but its transformative power has unveiled a paradox: the very vulnerabilities that could undermine our digital society are also the catalysts for a new, more resilient era. As the incidents from data poisoning to deepfake scams prove, the risks posed by AI are fundamentally different from those of traditional cybersecurity, and they can inflict real-world damage at an unprecedented scale.

To navigate this new reality, a global consensus is rapidly forming. **Standards like ISO/IEC 42001 and NIST's AI Risk Management Framework are providing the blueprints for a structured defense.** Simultaneously, national strategies, from the U.S. push for innovation to India's emphasis on data sovereignty, are converging on a shared objective to foster AI that is both powerful and secure.

The path forward demands decisive action. This isn't about slowing innovation; it's about embedding resilience into its core. By operationalizing security through AI Bills of Materials, rigorous red teaming, and robust data governance, we can move from reactive defense to proactive protection. Our commitment to these principles will be the ultimate determinant of whether AI strengthens our societies or leaves them vulnerable. By acting with foresight now, we ensure that AI's promise is realized, not at the expense of trust, but because of it.