Managing AI Security Risks

Creating an AI Security Framework

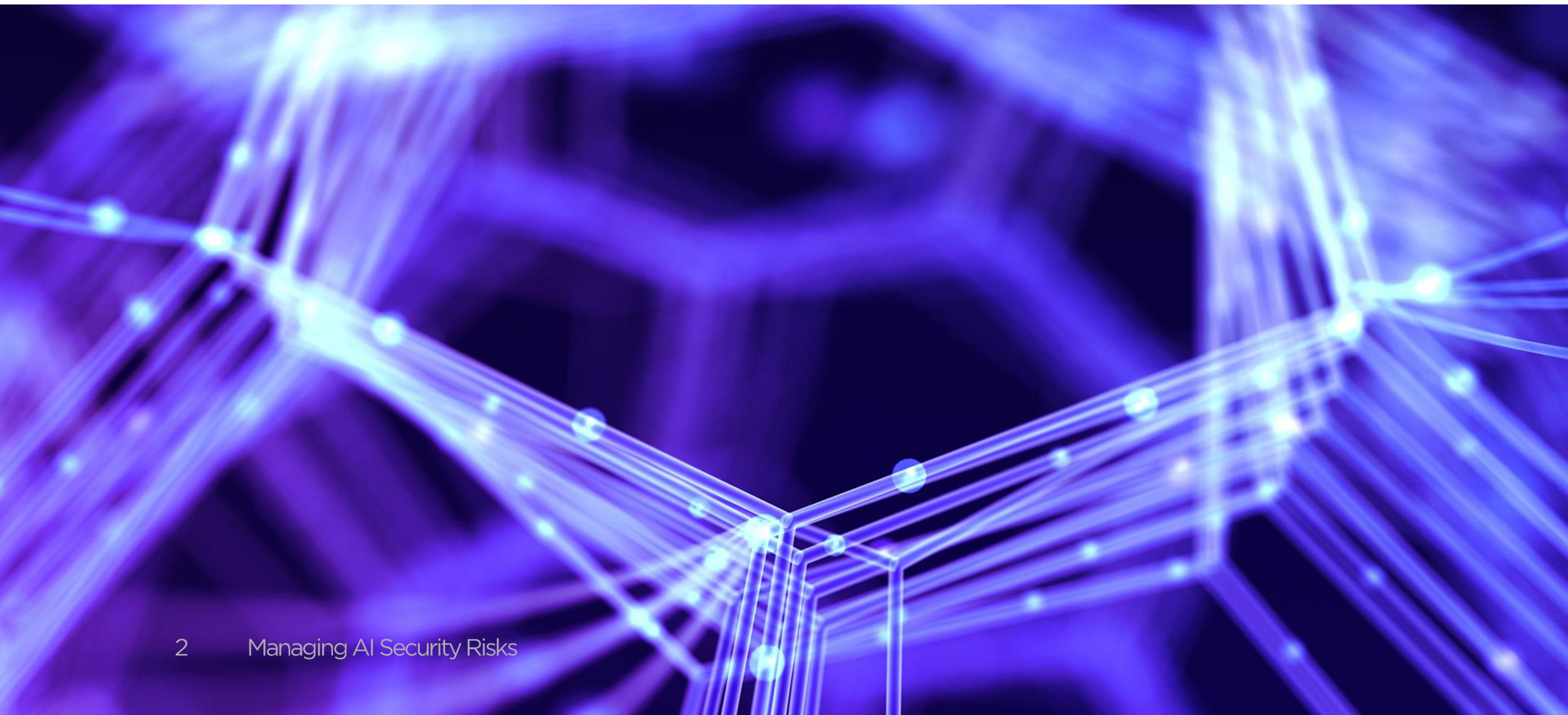
Executive Summary

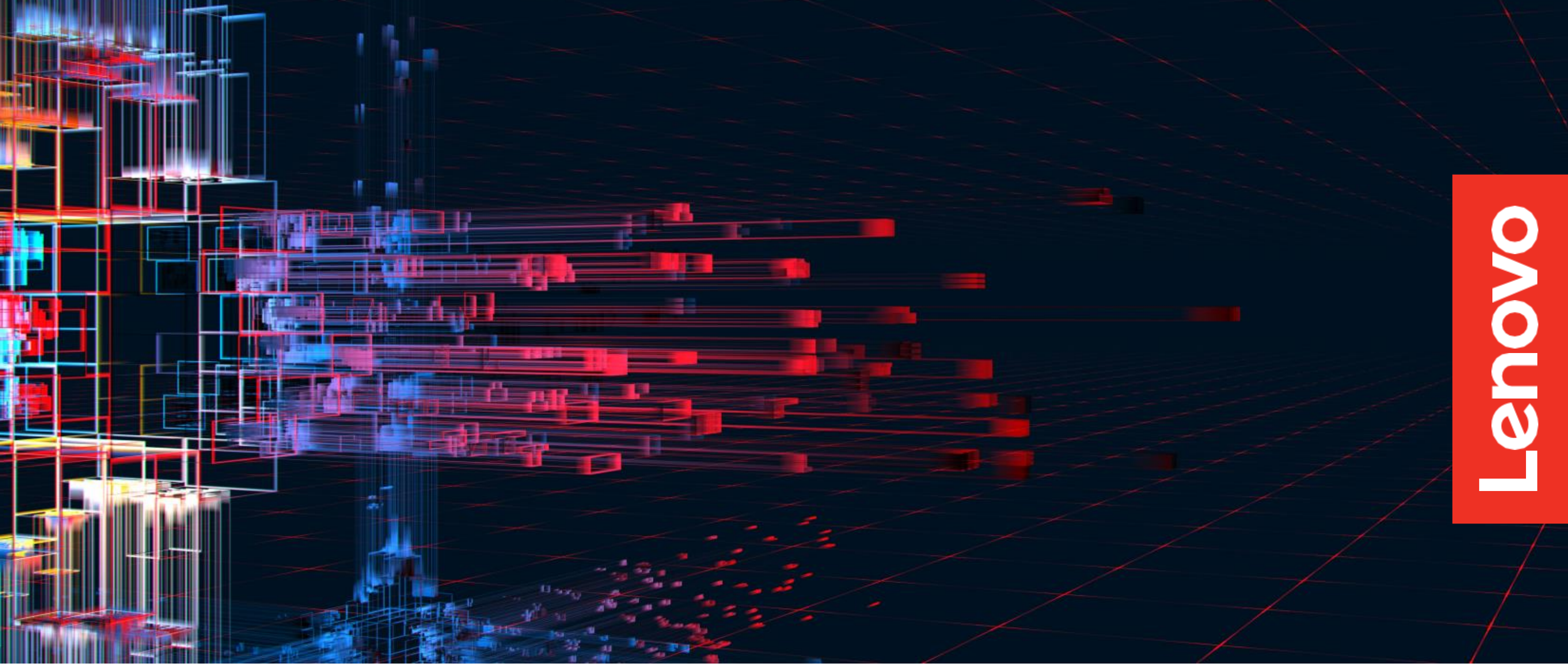
In today's rapidly evolving digital landscape, artificial intelligence (AI) has become a critical driver of innovation and operational efficiency. However, the deployment of AI technologies across organizations also presents significant security challenges. The increased complexity of AI systems, coupled with their expansive reach across the enterprise, has created new vulnerabilities that extend beyond traditional cybersecurity concerns and requires a broader aperture to manage. According to research, AI-enhanced malicious attacks, including AI-assisted misinformation have ranked as the top emerging risks for 2024[1]. As the accessibility and sophistication of AI tools increase, so does the potential for more devastating attacks. This introduces not only cybersecurity threats, but also security challenges related to model safety, data protection, and privacy, making it harder for organizations to respond effectively[2].

A recent Deloitte study further reveals that three of the top four barriers to the successful development and deployment of Generative AI (Gen AI) tools and applications are linked to security challenges. These include regulatory compliance (cited by 36% of respondents), difficulty managing risks (30% of respondents), and the lack of a governance model (29% of respondents)[3]. This is largely due to the legal uncertainties surrounding AI adoption, including evolving privacy regulations such as General Data Protection Regulation (GDPR), the European Union AI Act (EU AI Act), and the California Consumer Privacy Act (CCPA), vertical specific risks (higher scrutiny and complexity in industries like healthcare and financial services), as well as rising consumer expectations for transparent and fair AI systems.

To be truly secure, organizations must not only protect themselves from external cyber threats but also address internal risks, business vulnerabilities, and legal and reputational risks. A secure organization is one that defends itself from all these risk vectors—whether cyber, business, or regulatory in nature—ensuring robust defenses across the enterprise. While AI introduces security risks, the business benefits—such as operational efficiency, automation, and innovation—far outweigh these challenges when the risks are managed intentionally. By adopting AI, organizations can gain a competitive edge with data-driven insights and streamlined processes. The risks can be effectively managed through a robust AI Security Risk Framework that ensures data protection, model safety, and compliance. With the right security measures, the value of AI adoption becomes a strategic necessity. This requires building organizational capacity for proactive risk management and establishing robust security protocols that align with both regulatory requirements and emerging threats.

This paper outlines the security risks and vulnerabilities inherent in organizations AI systems, along with a comprehensive AI Risk Framework. It offers suggestions for organizations





across process and technology capabilities to mitigate these challenges, strategic priorities for leadership, and actionable steps for IT Decision Makers (ITDMs) and Chief Information Security Officers (CISOs) along with other executives to more resiliently navigate AI's complex security landscape.

Identifying and Mitigating AI Security-Related Risks

The integration of AI across enterprises has not only redefined value creation but has also simultaneously exposed organizations to sophisticated security risks. Across the various stages of the AI lifecycle, security risks are pervasive. For example, poor data quality or biased datasets during **data acquisition** can lead to inaccurate or unfair model outputs, while improper data handling may result in non-compliance with privacy regulations like the EU AI Act, GDPR or CCPA. During **model development**, issues such as bias, lack of explain-ability, and adversarial attacks like data poisoning threaten model integrity[4]. Insufficient testing in the **deployment** phase can introduce operational vulnerabilities, and integrating AI with legacy systems may create security loopholes. Failure to detect model drift during **evaluation** can lead to outdated or inaccurate predictions while inadequate **monitoring** might leave AI systems exposed to threats such as model inversion attacks or unauthorized access.

However, viewing these risks solely through the lens of the AI lifecycle can be limiting. AI security threats are complex, with risk vectors spreading across technical, regulatory, and operational domains. Addressing these risks in isolation—whether focusing on cyber threats or Governance, Risk, and Compliance (GRC) frameworks—fails to account for their interconnected nature[5]. Addressing both holistically allows organizations to manage technical vulnerabilities while maintaining alignment with regulatory standards, ethical guidelines, and broader risk management strategies.

This approach ensures that security and governance efforts complement each other so that an organization is resilient against a wider spectrum of threats—both internal and external. The '***AI Security Risk Framework***' provides organizations with a robust methodology for managing diverse AI security risks. This framework identifies six key dimensions—Model, Data, Malicious Use, Financial and Business, Regulatory and Legal, and Brand and Trust. The technical breadth of this framework allows organizations to develop targeted defense mechanisms that mitigate security threats ranging from data integrity violations to reputational harm. These security risk dimensions include:

- **Model Risks:** Focuses on technical vulnerabilities in AI models, addressing issues such as explain-ability, bias detection, and transparency to ensure systems remain accountable and fair. In addition to existing concerns, model integrity is a significant risk. Damaged or compromised model parameters can lead to inaccurate or unintended outputs, which are often difficult to identify. This poses a critical threat to the reliability and trustworthiness of AI systems.

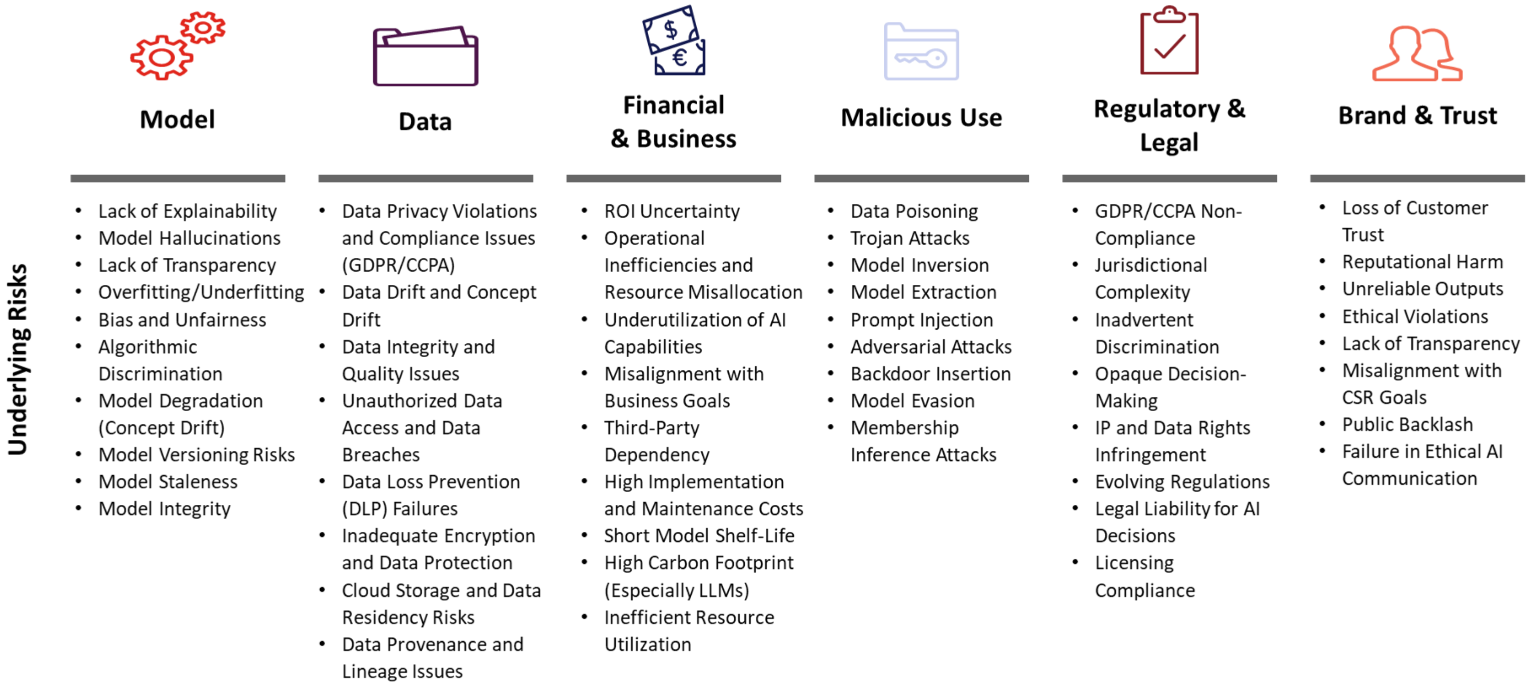
- **Data Risks:** Focuses on ensuring data integrity, privacy, and security across the AI lifecycle. This includes compliance with regulations like GDPR and CCPA, while addressing issues like data drift and poor-quality inputs that affect model

performance. It also includes risks such as data breaches and encryption failures.

- **Malicious Use Risks:** Targets defense against adversarial attacks such as data poisoning, model inversion, and prompt injection, which, if compromised, can lead to significant operational disruption.

- **Financial and Business Risks:** Explores the economic sustainability of AI, covering operational costs, return on investment (ROI), and environmental impacts such as energy consumption and the carbon footprint of data centers. It also aligns AI operations with broader organizational goals for profitability and sustainability.

- **Regulatory and Legal Risks:** Addresses compliance with evolving legal landscapes, including adherence to regulations like GDPR, CCPA, and sector-specific legislation (such as more stringent regulations and privacy considerations Healthcare and Financial Services), while tackling intellectual property, liability, and AI ownership issues.



Underlying Risks

Model

- Lack of Explainability
- Model Hallucinations
- Lack of Transparency
- Overfitting/Underfitting
- Bias and Unfairness
- Algorithmic Discrimination
- Model Degradation (Concept Drift)
- Model Versioning Risks
- Model Staleness
- Model Integrity

Data

- Data Privacy Violations and Compliance Issues (GDPR/CCPA)
- Data Drift and Concept Drift
- Data Integrity and Quality Issues
- Unauthorized Data Access and Data Breaches
- Data Loss Prevention (DLP) Failures
- Inadequate Encryption and Data Protection
- Cloud Storage and Data Residency Risks
- Data Provenance and Lineage Issues

Financial & Business

- ROI Uncertainty
- Operational Inefficiencies and Resource Misallocation
- Underutilization of AI Capabilities
- Misalignment with Business Goals
- Third-Party Dependency
- High Implementation and Maintenance Costs
- Short Model Shelf-Life
- High Carbon Footprint (Especially LLMs)
- Inefficient Resource Utilization

Malicious Use

- Data Poisoning
- Trojan Attacks
- Model Inversion
- Model Extraction
- Prompt Injection
- Adversarial Attacks
- Backdoor Insertion
- Model Evasion
- Membership Inference Attacks

Regulatory & Legal

- GDPR/CCPA Non-Compliance
- Jurisdictional Complexity
- Inadvertent Discrimination
- Opaque Decision-Making
- IP and Data Rights Infringement
- Evolving Regulations
- Legal Liability for AI Decisions
- Licensing Compliance

Brand & Trust

- Loss of Customer Trust
- Reputational Harm
- Unreliable Outputs
- Ethical Violations
- Lack of Transparency
- Misalignment with CSR Goals
- Public Backlash
- Failure in Ethical AI Communication

Figure 1a: AI Security Risk Framework

Strategies for Holistically Mitigating AI Security Risks

Effectively managing AI-related security risks requires a combination of general and highly targeted mitigation strategies. Universal approaches, such as continuous model monitoring through automated performance metrics and anomaly detection, and bias auditing via fairness assessments, are crucial for maintaining the integrity of AI systems. Stringent data governance frameworks that enforce data lineage tracking, version control, and access management further safeguard the system's resilience. Select tangible recommendations ITDMs can act on quickly are captured:

- Adopting cloud-native backup solutions with end-to-end encryption and data redundancy across distributed architectures to provide comprehensive protection against breaches[6].
- Deploy advanced adversarial training to harden models against manipulation
- Leverage Intrusion Detection Systems (IDS)[7] to monitor for unusual access patterns or attack vectors
- Routinely revalidate models through retraining pipelines to combat model degradation or concept drift
- Enforce comprehensive AI governance frameworks that automate compliance checks and audit trails for regulatory standards (e.g., GDPR, CCPA)
- Enforce robust endpoint security to protect AI systems accessed via endpoint devices, ensuring comprehensive protection against unauthorized access and potential compromises. This includes the use of firewalls, antivirus software, and encryption to secure data and models at vulnerable entry points.

The following figure further expands on mitigation strategies mapped to each risk dimension:

Risk Mitigation Strategies

Model

- **Lack of Explainability:** Use XAI tools like LIME, SHAP, or integrated gradients to improve model transparency and meet regulatory requirements
- **Model Hallucinations**: Apply constrained decoding and fine-tune with high-quality datasets. Implement output verification to ensure accuracy
- **Lack of Transparency:** Enforce audit trails and version control systems (e.g., Git, MLflow) to ensure traceability in model development
- **Overfitting / Underfitting:** Use cross-validation, regularization (L1, L2), and early stopping to optimize model generalization and avoid overfitting
- **Bias and Unfairness:** Apply fairness auditing frameworks (e.g., AI Fairness 360) and retrain models with diverse datasets. Use adversarial debiasing for bias mitigation
- **Algorithmic Discrimination:** Conduct regular bias audits and adversarial testing to assess and mitigate discriminatory impacts
- **Model Degradation (Concept Drift):** Monitor model performance using drift detection algorithms (e.g., KL divergence) and implement active learning pipelines to keep models updated
- **Model Versioning Risks:** Use automated versioning tools (e.g., MLflow, DVC) to maintain consistency and ensure reproducibility across environments
- **Model Staleness:** Schedule routine retraining and use lifecycle management tools to refresh models with new data as needed
- **Model Integrity**: Implement hardware-based protections like Trusted Execution Environments (TEEs) to secure model parameters and prevent tampering, ensuring reliable AI outputs.

Data

- **Data Privacy Violations and Compliance Issues (GDPR/CCPA):** Enforce AES-256 and homomorphic encryption for secure analysis. Use differential privacy and privacy-by-design to ensure regulatory compliance
- **Data Drift and Concept Drift:** Implement data versioning to maintain historical integrity and enable continuous drift monitoring. Use automated tools to detect distribution changes early
- **Data Integrity and Quality Issues:** Use validation tools to enforce quality controls. Apply real-time data masking and sandboxing to protect sensitive information during analysis
- **Unauthorized Data Access and Data Breaches:** Employ role-based access control (RBAC), multi-factor authentication (MFA), and intrusion detection systems (IDS) for monitoring and limiting access to sensitive data
- **Data Loss Prevention (DLP) Failures:** Leverage DLP solutions that include real-time content inspection and contextual data analysis to prevent unauthorized data leakage across hybrid and cloud environments
- **Inadequate Encryption and Data Protection:** Enforce end-to-end encryption across on-premises and cloud environments. Implement real-time data masking and tokenization for sensitive data, ensuring privacy without hindering analysis
- **Cloud Storage and Data Residency Risks:** Use Hybrid AI to store sensitive data on-premises while leveraging cloud for scalable processing. Ensure compliance with residency laws using geo-redundant cloud backups
- **Data Provenance and Lineage Issues:** Implement data lineage tracking and metadata management systems to maintain transparency over the data lifecycle and ensure traceability from source to destination
- **Endpoint Security Vulnerabilities:** Secure endpoint devices with firewalls, encryption, and intrusion detection to prevent unauthorized access to data and models on AI systems

Risk Mitigation Strategies

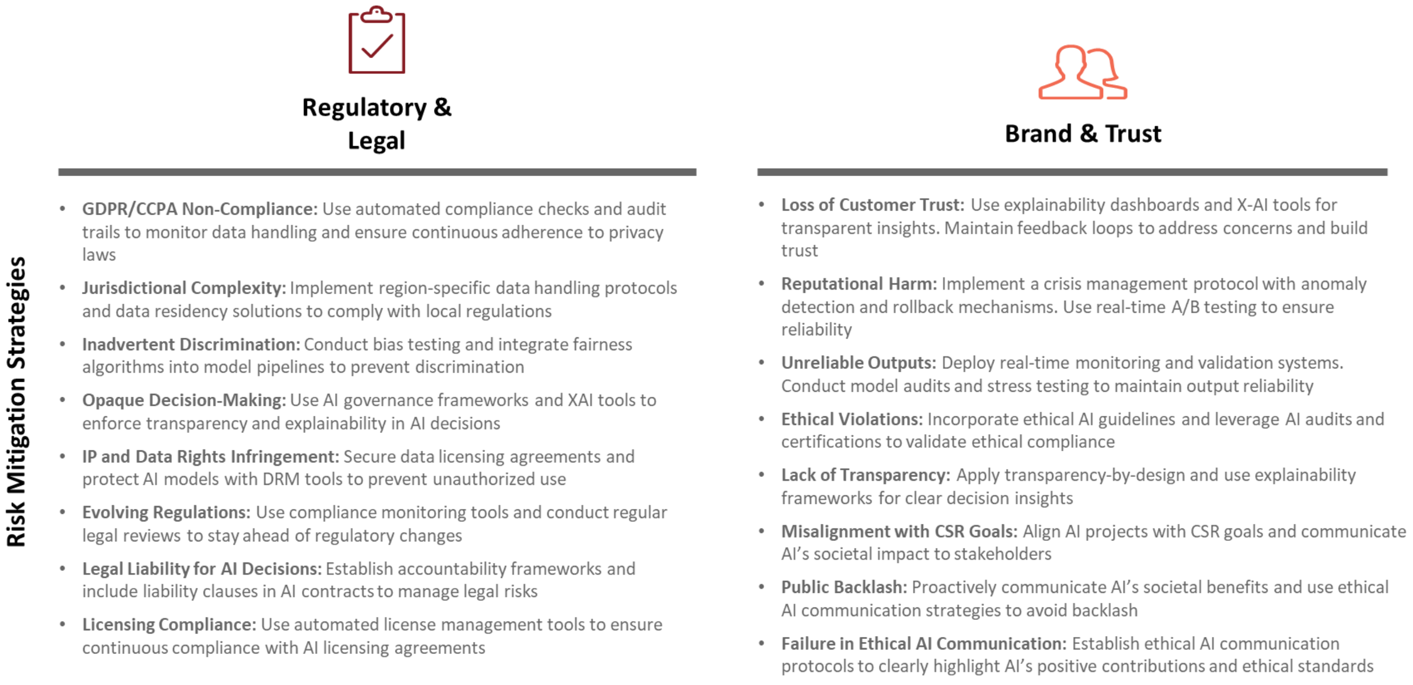
Financial & Business

- **ROI Uncertainty:** Use automated financial models and well-defined KPIs to track ROI. Conduct regular cost-benefit analyses to ensure AI projects align with business goals
- **Operational Inefficiencies**: Implement scalable cloud infrastructure and containerization to dynamically allocate resources and reduce inefficiencies
- **Underutilization of AI Capabilities:** Regularly reassess AI utility and integrate feedback from business units to maximize AI usage
- **Misalignment with Business Goals:** Align AI initiatives with business strategies using real-time performance dashboards to ensure they drive business outcomes
- **Third-Party Dependency:** Diversify AI vendors, implement contractual safeguards, and explore in-house AI development to reduce vendor reliance
- **High Implementation Costs:** Use predictive cost management tools and cloud-based solutions to minimize infrastructure and maintenance costs
- **Short Model Shelf-Life:** Establish continuous retraining pipelines and model lifecycle management to ensure timely updates and extend model utility
- **High Carbon Footprint (LLMs):** Leverage liquid cooling systems to optimize energy usage in data centers. Adopt energy-efficient hardware, use renewable energy sources, and optimize training cycles to lower environmental impact
- **Inefficient Resource Utilization:** Adopt a hybrid environment (cloud, on-premises, data center) to balance resource utilization. Use virtualization and containerization across these environments for optimized resource allocation and efficiency

Malicious Use

- **Data Poisoning:** Implement robust input validation and filtering mechanisms to defend against malicious data. Use outlier detection to block poisoned data
- **Trojan Attacks:** Conduct regular penetration testing and apply security patches. Use code auditing to detect trojans before they exploit model vulnerabilities
- **Model Inversion:** Utilize federated learning and privacy-preserving techniques like differential privacy to reduce model exposure and prevent inversion attacks
- **Model Extraction:** Limit exposure by using watermarking techniques and restricting access with rate-limiting. Apply privacy-preserving techniques like homomorphic encryption to protect sensitive data
- **Prompt Injection:** Implement input validation and real-time filtering for malicious prompts. Continuously monitor prompt behavior and apply Zero Trust Architecture (ZTA) for user verification
- **Adversarial Attacks:** Use adversarial training to increase robustness. Conduct regular security audits and implement defensive distillation to harden models against small perturbations
- **Backdoor Insertion:** Perform model integrity checks and code analysis. Retrain models with verified data and regularly audit the model pipeline for hidden backdoors
- **Model Evasion:** Apply adaptive security mechanisms to detect and respond to evasion attempts. Use dynamic model updates to adjust models as evasion techniques evolve
- **Membership Inference Attacks:** Leverage differential privacy and noise injection to obscure sensitive data and reduce the risk of inference attacks

Figure 1b: AI Security Risk Mitigation Strategies



Risk Mitigation Strategies

Regulatory & Legal

- **GDPR/CCPA Non-Compliance:** Use automated compliance checks and audit trails to monitor data handling and ensure continuous adherence to privacy laws
- **Jurisdictional Complexity:** Implement region-specific data handling protocols and data residency solutions to comply with local regulations
- **Inadvertent Discrimination:** Conduct bias testing and integrate fairness algorithms into model pipelines to prevent discrimination
- **Opaque Decision-Making:** Use AI governance frameworks and XAI tools to enforce transparency and explainability in AI decisions
- **IP and Data Rights Infringement:** Secure data licensing agreements and protect AI models with DRM tools to prevent unauthorized use
- **Evolving Regulations:** Use compliance monitoring tools and conduct regular legal reviews to stay ahead of regulatory changes
- **Legal Liability for AI Decisions:** Establish accountability frameworks and include liability clauses in AI contracts to manage legal risks
- **Licensing Compliance:** Use automated license management tools to ensure continuous compliance with AI licensing agreements

Brand & Trust

- **Loss of Customer Trust:** Use explainability dashboards and X-AI tools for transparent insights. Maintain feedback loops to address concerns and build trust
- **Reputational Harm:** Implement a crisis management protocol with anomaly detection and rollback mechanisms. Use real-time A/B testing to ensure reliability
- **Unreliable Outputs:** Deploy real-time monitoring and validation systems. Conduct model audits and stress testing to maintain output reliability
- **Ethical Violations:** Incorporate ethical AI guidelines and leverage AI audits and certifications to validate ethical compliance
- **Lack of Transparency:** Apply transparency-by-design and use explainability frameworks for clear decision insights
- **Misalignment with CSR Goals:** Align AI projects with CSR goals and communicate AI's societal impact to stakeholders
- **Public Backlash:** Proactively communicate AI's societal benefits and use ethical AI communication strategies to avoid backlash
- **Failure in Ethical AI Communication:** Establish ethical AI communication protocols to clearly highlight AI's positive contributions and ethical standards

Figure 1b: AI Security Risk Mitigation Strategies

Adopting these strategies can help mitigate risks so that organizations can maximize the value of their AI investments. A robust security posture helps ensure that organizations remain on their front foot so that they can capitalize on the full potential of AI while decreasing exposure to internal and external threats.

Building Process Capabilities to Strengthen AI Security

Strengthening AI security processes requires a shift from reactive to proactive and predictive strategies. A reactive approach often results in delayed responses to emerging threats, leaving AI systems vulnerable to breaches and attacks. By embedding security at every stage of the AI lifecycle, organizations can reduce attack surfaces, anticipate potential vulnerabilities, and respond faster to evolving threats.

1. **Shift from Reactive to Proactive and Predictive Security:** To manage AI security risks more proactively, organizations should consider:

- **Advanced Security Tools with Development, Security, and Operations (DevSecOps):** Incorporate automated security tools directly into the Continuous Integration and Continuous Deployment (CI/CD) pipeline, ensuring continuous vulnerability assessment throughout the AI lifecycle. Tools should monitor model integrity, data security, and infrastructure security during deployment to enable early detection of security flaws and reduces exposure to attack vectors.

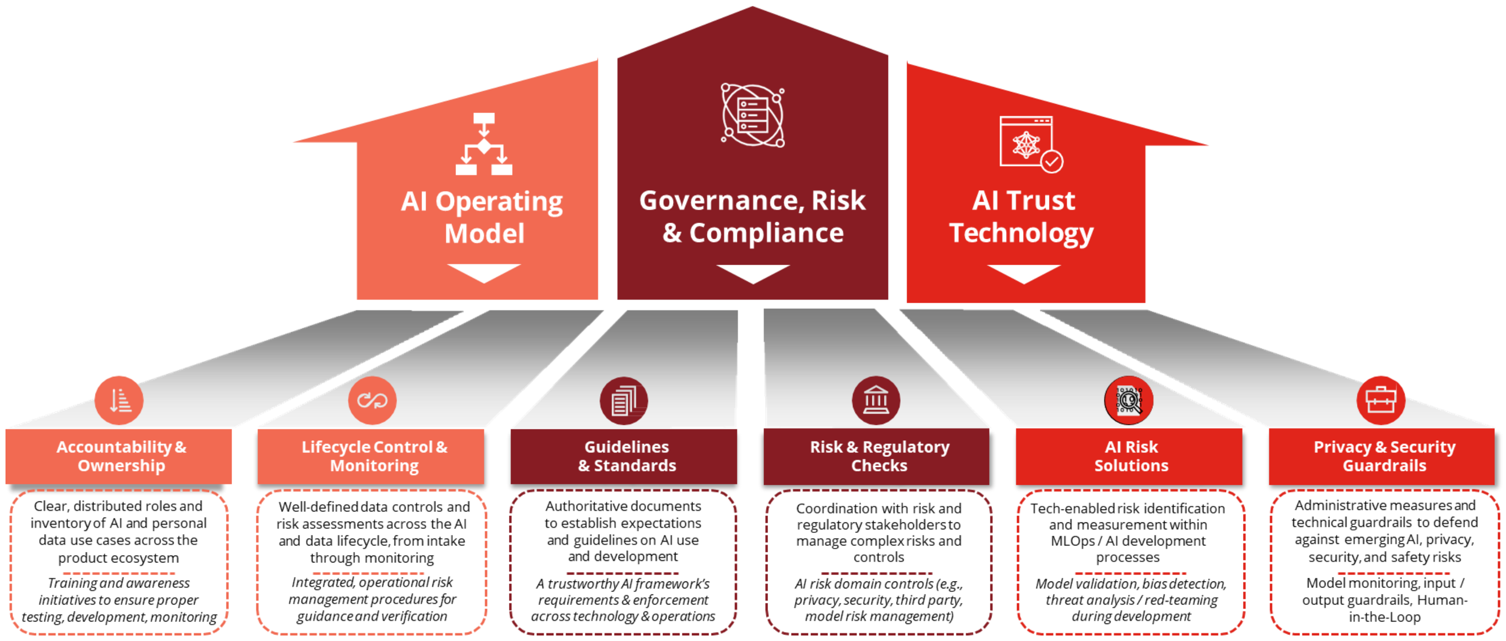
- **Model Updates through Large Language Model Operations (LLMOps):** Use LLMOps to automate model maintenance and patching, ensuring that models receive timely updates to stay secure[8]. Continuous retraining pipelines should be set up to prevent model drift and ensure up-to-date threat protection. This enhances model security without manual intervention.

> **"A robust AI Security Risk Management framework helps ensure that AI systems operate reliably, ethically, and in compliance with regulatory standards."**

- **Red Teaming and Penetration Testing:** Regularly conduct red teaming exercises and penetration tests to simulate cyberattacks on AI models, infrastructure, and Application Programming Interfaces (APIs). These tests should mimic adversarial threats, backdoor attacks, and data exfiltration attempts.
- **Automated Incident Response:** Integrate automated incident response systems that can detect anomalies, execute pre-defined remediation actions, and rollback models or systems when threats are detected. Ensure critical actions are flagged for human review to maintain accountability and prevent unintended impacts. Combine with AI-driven Security Operations Centre (SOC) systems for real-time incident response. This reduces potential downtime through rapid threat containment.
- **Threat Intelligence Powered by Zero Trust Architecture:** Leverage threat intelligence platforms (TIPs) for real-time tracking of emerging vulnerabilities and risks. Use Zero Trust Architecture (ZTA) to enforce continuous identity verification, micro-segmentation of networks, and strict access control policies. This ensures that only verified users access AI systems.

2. **Establish a Robust AI Security Risk Management Framework:** A robust AI Security Risk Management framework helps ensure that AI systems operate reliably, ethically, and in compliance with regulatory standards.

- It’s three pillars include the AI Operating Model, GRC, and AI Trust Technology[9]. The AI operating model provides a structured approach for managing the AI lifecycle, aligned to business objectives and operational processes. GRC ensures that AI systems adhere to legal requirements, mitigate security risks, and maintain organizational integrity. AI Trust Technology encompasses the tools and techniques that enhance the transparency, fairness, and accountability of AI models. Together, these pillars form a comprehensive security risk management framework that safeguards AI security and fosters trust.



AI Operating Model
Governance, Risk & Compliance
AI Trust Technology
Accountability & Ownership
Clear, distributed roles and inventory of AI and personal data use cases across the product ecosystem
Training and awareness initiatives to ensure proper testing, development, monitoring
Lifecycle Control & Monitoring
Well-defined data controls and risk assessments across the AI and data lifecycle, from intake through monitoring
Integrated, operational risk management procedures for guidance and verification
Guidelines & Standards
Authoritative documents to establish expectations and guidelines on AI use and development
A trustworthy AI framework’s requirements & enforcement across technology & operations
Risk & Regulatory Checks
Coordination with risk and regulatory stakeholders to manage complex risks and controls
AI risk domain controls (e.g., privacy, security, third party, model risk management)
AI Risk Solutions
Tech-enabled risk identification and measurement within MLOps / AI development processes
Model validation, bias detection, threat analysis / red-teaming during development
Privacy & Security Guardrails
Administrative measures and technical guardrails to defend against emerging AI, privacy, security, and safety risks
Model monitoring, input / output guardrails, Human-in-the-Loop

Figure 2: AI Security Risk Management Framework

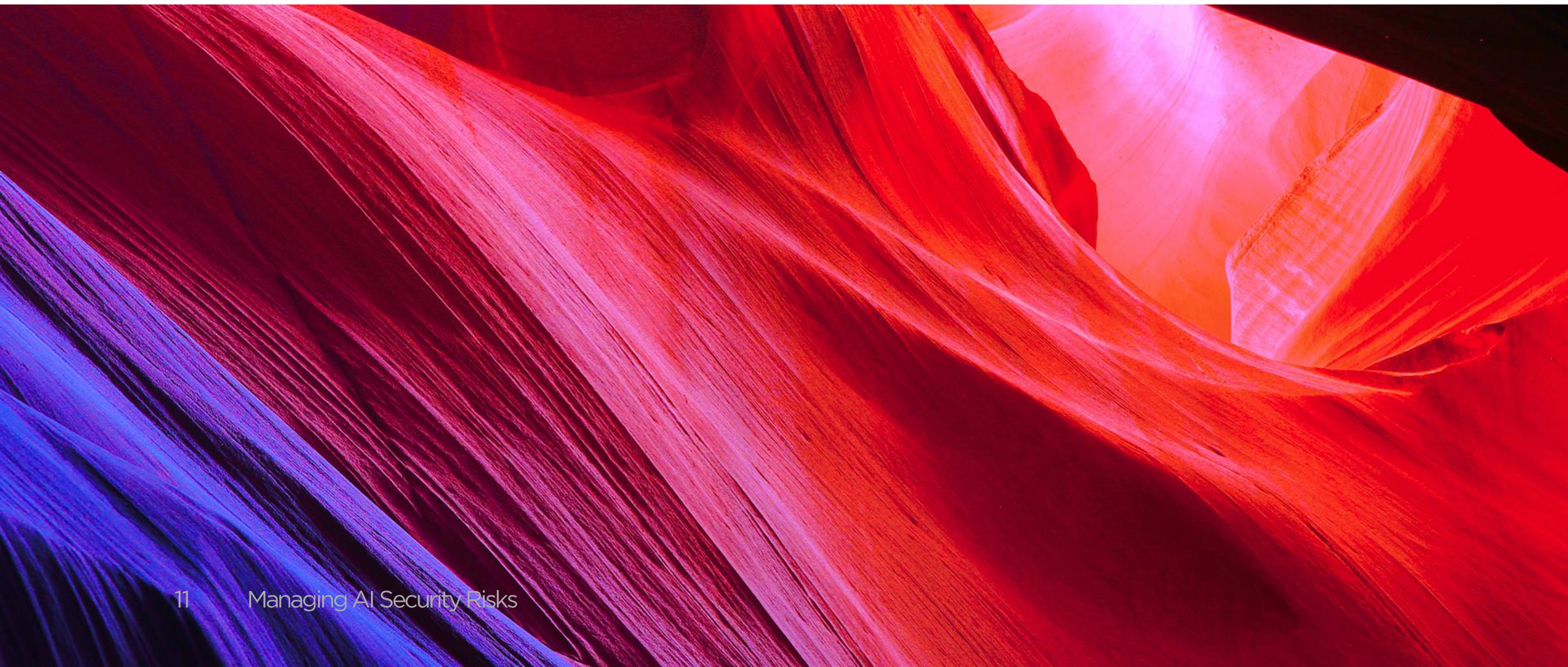
- **AI Operating Model with Security Controls and Accountability:** Implement a structured AI operating model that integrates security into each phase of the AI lifecycle, from model design to deployment and monitoring. Establish accountability and ownership tracking mechanisms to clearly define roles and responsibilities for managing AI risks. Use security Key Performance Indicators (KPIs) to ensure AI systems meet organizational and regulatory standards.

- **GRC with Monitoring Guidelines:** Use a GRC platform to continuously track and assess compliance with regulatory frameworks such as GDPR, CCPA, and industry-specific standards. Define and implement monitoring guidelines and standards to enforce consistent oversight of AI systems. Perform automated compliance audits for real-time visibility into legal and regulatory adherence.

- **AI Trust Technology with Risk Solutions and Privacy/Security Guardrails:** Deploy AI risk management solutions that enhance the transparency, fairness, and accountability of AI models. Employ Explainable AI (XAI) tools, privacy-preserving techniques (such as differential privacy and homomorphic encryption), and security guidelines to ensure AI models are aligned with ethical and privacy standards. Implement privacy and security guidelines to govern data use and model operations.

3. **Standardize Monitoring and Auditing:** To effectively detect and mitigate AI security breaches[10] organizations should consider.

- **Centralized Monitoring through Security Information and Event Management (SIEM):** Deploy SIEM systems to aggregate security event logs from AI models, data infrastructure, and operational tools. Implement real-time telemetry across the infrastructure for continuous threat detection. These actions provide centralized visibility and early detection of security anomalies across the AI ecosystem.
- **Behavioral Analytics and Intrusion Detection:** Implement behavioral analytics and Intrusion Detection Systems (IDS) to monitor network traffic for suspicious patterns, identify abnormal behaviors including insider threats, and flag security incidents. Integrate machine learning models to predict and prevent unauthorized access.



- **LLMOps for Continuous AI Monitoring:** Utilize LLMOps to continuously monitor AI models for performance degradation, model drift, and data drift. Implement model revalidation pipelines to ensure models maintain accuracy and integrity.
- **Integration with Incident Response and SOC:** Connect monitoring and auditing systems with SOC to automate incident response based on predefined playbooks. Use real-time analytics and AI-driven incident detection to improve response speed and remediation of security incidents.
- **Post-Incident Forensics and Auditing:** After a security incident, conduct forensic analysis and post-incident reviews to understand the root cause, update security control and reduce repeat incidents. Audit model and data handling practices to ensure no repeat vulnerabilities.

About Lenovo

References

[1] Gartner: AI-Enhanced Malicious Attacks Are a New Top Emerging Risk for Enterprises

[2] Forrester: Securing Generative AI

[3] Deloitte: Now Decides Next

[4] TechTarget: A guide to artificial intelligence in the enterprise

[5] Forrester: NIST AI Risk Management Framework 1.0

[6] Lenovo: Predict, Prevent, Protect

[7] Gartner: Tackling Trust, Risk and Security in AI Models

[8] Forrester: Defending AI And Generative AI Models

[9] Deloitte: AI Risk Management

[10] TechTarget: A guide to artificial intelligence in the enterprise