Building Blocks of Secure AI Governance Framework

New York Metro Joint Cyber Security Conference - 2024

AI GOVERNANCE IS CRITICAL

We understand trustworthy and ethical AI is a complex business, regulatory, and technical challenge, and we are committed to helping Clients put into practice. We help developed, and deploy an end-to-end trusted AI program across the AI/ML life cycle

Fairness
Fairness ensure model reduce or eliminates biased against individuals, communities or groups

Privacy
Ensure compliance with Data privacy regulations and consumer data usage.

Transparency
Include responsible disclosure to provide stakeholders a clear understanding as to what is happening within the AI solution and across the AI lifecycle

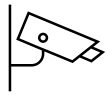
Sustainability
Optimize AI solutions to limit negative environmental impact where possible

Explainability
Ensure AI solutions are under stable as to how and why recommendations are made, or conclusions drawn

Data Integrity
Ensure data quality, governance and enrichment steps embedded trust

Accountability
Human oversight and responsibility embedded across the AI lifecycle to manage risk and ensure compliance with the regulations and applicable laws

Reliability
Ensure AI systems performs at the desired level of precession and consistency



Security
Safeguard against unauthorized access, bad actors, misinformation, corruption, or attacks

Safety
Safeguard AI solutions against harm to human and property

EXL

EVER REVOLVING REGULATORY ENVIRONMENT

Core Governance Principle	Fairness	Explainability	Integrity of data	Security and Resiliency	Accountability	Privacy	Risk approach
Desc of principles	Fair and equitable outcomes across different groups	Ability to explain how AI outcomes are achieved	Leverage High quality appropriate data will Lineage	Design AI to operate as intended with security	Human responsibility for AI decision outcomes	Respect and protect right of consumer data	Targeted risk identification and assessment
Global Regulatory guidance							
National AI Initiative act	☑	☑	☑	☑		☑	☑
AI in government	☑	☑	☑	☑	☑		
The National AI resource tasks force	☑ ☑	☑ ☑	☑	☑ ☑	☑ ☑	☑ ☑	☑
NIST AI Risk framework	☑	☑	☑	☑	☑	☑	☑
FHFAAB 2020-02	☑			☑	☑	☑	☑
NAIC Principle on AI	☑	☑			☑		
Federal trade commission	☑	☑		☑	☑	☑	
EU AI act	☑	☑		☑	☑	☑	☑
EU Digital service act	☑	☑			☑	☑	☑
OECD Principles							

RISK MANAGEMENT APPROACH FOR AI – FOCUS ON CYBER AND DATA PROTECTION

- Evolving regulations, client demands and alignment to industry best practices are playing a pivotal role in determining AI security roadmaps
- With risk-based lens, concerted efforts are being made to progressively align cyber practices and heighten security assurance for AI enabled solutions
- Secure AI program shall focus on strengthening governance, data protection, regulatory adherence and upskill workforce on emerging technologies

Governance

- AI Governance Committee for oversight on high-risk AI models; ongoing risk management by AI Ops Committee, Cyber and Business Unit

Framework and Processes

- Right AI Risk Management Framework, Responsible AI Governance Policy, AI Acceptable Usage Guidelines, other applicable cyber processes*

Tools and Technology

A. Existing cyber security capabilities to safeguard risks pertaining to data leakage, access mgmt., cyber threats, legal and regulatory disclosures
B. Deploying and progressively enhancing advanced AI risk mgmt. toolset for privacy threat visualization and mitigation

Awareness and Culture

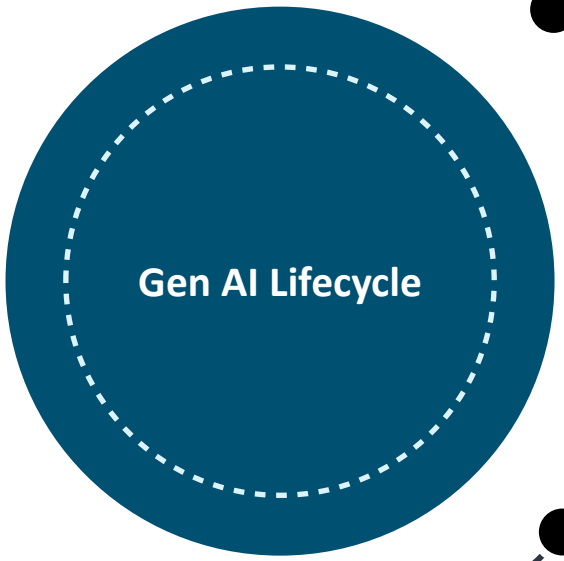
- Driving multi-channel awareness campaigns (Workshops, Webinars, Mailers), cyber simulation exercises covering AI solutions

Monitoring and Compliance

- Threat intel and monitoring by SOC [2], incident and breach response playbooks, specialized partner ecosystem (Forensics, Dark web monitoring etc.)

1: NIST - National Institute of Standards and Technology | 2. Security Operations Center * e.g., Application and Cloud Security, Data Privacy and Protection

RESPONSIBLE GENERATIVE AI LIFECYCLE



Data Collection & Preprocessing

It involves finding relevant and reliable sources of data and cleaning, labeling, augmenting, and splitting the data to ensure its quality, consistency, diversity

Model Selection

Process of choosing the best generative AI model for a given task, data set, and evaluation metric.

Model Training

Process of teaching a system to perform a specific task. It involves providing the system with data, and feedback, such as rewards, penalties

Model Evaluation

Measuring the performance and quality of a Gen AI model. It involves comparing the model's outputs with some reference data or criteria, such as metrics or benchmarks

Model Fine-Tuning

Adjusting the parameters of a pre-trained model to improve its performance on a specific task

Deployment & Maintenance

Process of making a trained Gen AI model available for use and ensuring its reliability and performance over time. It involves steps such as testing, monitoring and updating

SOME LEADING SECURE AI FRAMEWORKS

- NIST's Artificial Intelligence Risk Management breaks down AI security into four primary functions: govern, map, measure, and manage.
- AI RMF Generative AI Profile can help organizations identify unique risks posed by generative AI and proposes actions for generative AI risk management that best aligns with their goals and priorities.

- ISO/IEC 42001 is an international standard that specifies requirements for establishing, implementing, maintaining, and continually improving an Artificial Intelligence Management System (AIMS) within organizations
- ISO 42001 is designed for entities providing or utilizing AI-based products or services, ensuring responsible development and use of AI systems.

- MITRES Sensible Regulatory Framework for AI Security and ATLAS Matrix anatomize attack tactics and propose certain AI regulations.
- MITRE and Microsoft have collaborated to enhance the MITRE ATLAS™ (Adversarial Threat Landscape for Artificial-Intelligence Systems), which now includes a focus on generative AI vulnerabilities.

- Google SAIF (Secure AI Framework) is designed to provide a security framework or ecosystem for the development, use and protection of AI systems.
- SAIF is designed to help mitigate risks specific to AI systems, like stealing the model, data poisoning of the training data, injecting malicious inputs through prompt injection, and extracting confidential information in the training data.

SECURE AI GOVERNANCE CONSIDERATIONS AND RELEVANT CYBER CAPABILITIES

- Risk assurance approach is aligned with regulatory landscape and industry standards like – EU AI ACT, US NIST framework and ISO 42001 etc.
- Associated risks include privacy concerns, biased programming, unclear legal regulations, copyright issues and inherent security vulnerabilities
- With a risk-based lens, Cyber team is firming up a multi-pronged approach to progressively address the security concerns arising from increased AI exposure

Secure AI Governance Considerations

Relevant Capabilities

Data Sets
Relevant and approved data sources

- Privacy Threat Modelling (PII[1]/PFI[2]/PHI[3]) *
- Data Sensitivity Threat Visualization *
- Intelligent Recommendation

Data Security
Minimize data leakage probability

- Data Anonymization and Synthetic data *
- Pseudonymization and Differential Privacy
- Data Leakage Prevention (web and endpoint)

AI Bias
Eliminate biases in models

- Bias Detection and Measurement
- Fairness-aware Algorithm Design
- Diverse and Representative Data

Regulatory
Compliance with data regulations

- Privacy Impact Assessment
- Cross Border Data Transfer
- Privacy by Design

Cyber Threat
Strengthen cyber practices for AI risk

- Security Architecture and Code Reviews
- Web Application Firewall and API[4] Security
- Container Security and Configuration Hardening

Access Management
Restrict unauthorized data access

- Role-Based Access Control
- Multi Factor Authentication
- User and Entity Behavior Analytics (UEBA)

Consumer Privacy
Adherence to geo-specific obligations

- Data Subject Rights Management
- Consent Management
- Data Minimization

Licensing
Regulate legitimate software usage

- Secure Supply Chain
- Software Composition Analysis (SCA)
- AI model security

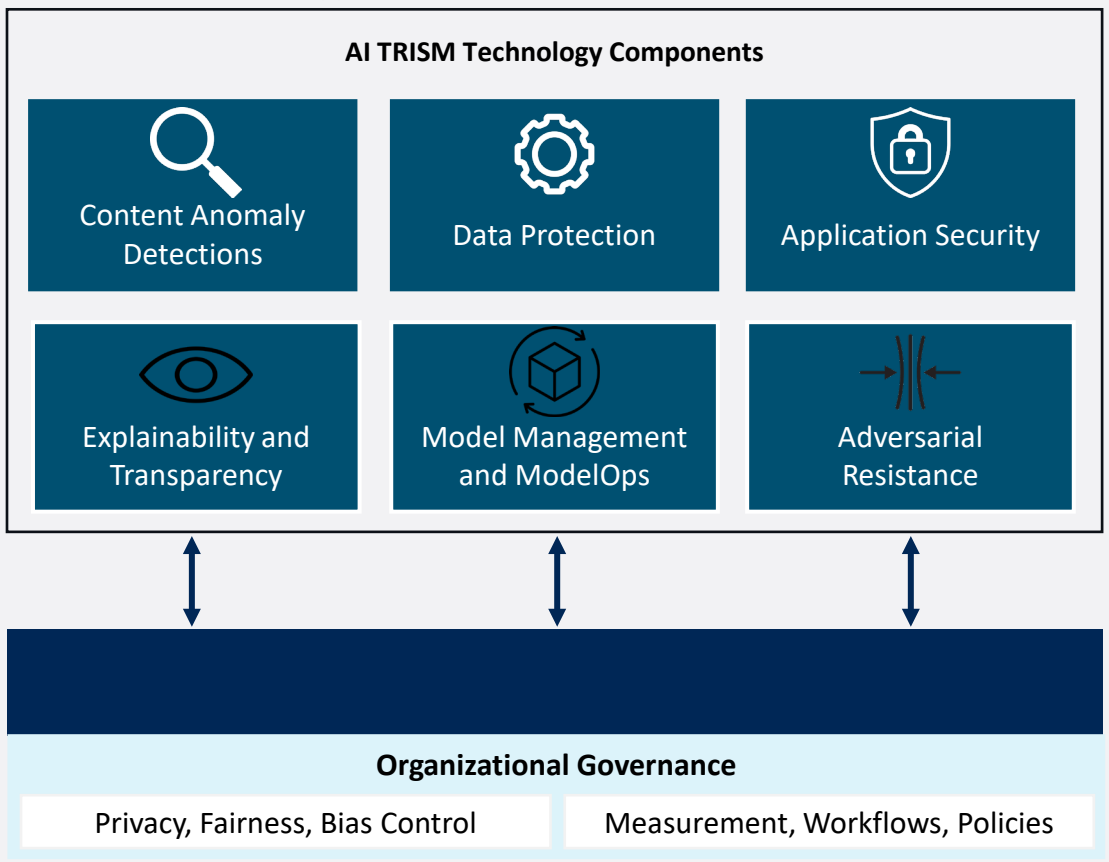
Solution assessment with privacy preserve technology assurance

United States	Canada	European Union	United Kingdom	India	Australia
AI Disclosure Act, California Executive Order on AI, Massachusetts Senate Bill	AI and Data Act (AIDA)	EU AI Act Article 28b	Existing Data Protection and Privacy laws apply	Digital Personal Data Protection Act	Privacy Act, Data Availability and Transparency Act

1: PII – Personal Identifiable Information | 2: PFI- Personal Financial Information | 3: PHI – Protected Health Information | 4: API – Application Programming Interface | 5: SIEM: Security Information and Event Management

KEY RESPONSIBLE AI PRINCIPLES

Responsible AI principles assurance as part of AI TRiSM framework



AI System users need to acquire this to fill gaps in builder/owner solutions

DATA PROTECTION

- Data Protection Impact Assessment
- Pseudonymization
- Privacy Risk Assessment
- Synthetic Data
- Anonymization
- Synthetic Prompts

APPLICATION SECURITY

- Role based Model Access
- Prompt Context Identification
- Model Usage Governance Report

ADVESARISAL RESISTANCE

- Direct Attacks
- Jail Break Attacks
- Harmful Usage
- Inappropriate Usage Attacks

MODELOPS

- Privacy Protected Data Preparation
- DPIA for Data access for model training
- Prediction Monitoring for Risks

CONTENT ANOMALY DETECTIONS

- Anomalous usage detection through prompt monitoring and prevention of certain kinds of attacks

The approach includes solutions for Data protection elements, Application Security for LLM usage, Adversarial resistance against modern LLM attacks, ModelOps support with Privacy protection and risk mitigation, along with Content anomaly detection in prompts.

1: TRiSM – AI Trust, Risk and Security Management by Gartner

Thank you

Rahul Bhardwaj
EXL
VP – Cyber and Global Privacy

EXL