AIShield Information for NIST “Related to NIST’s Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence”

Authors: Yuvaraj Govindarajulu, Manpreet Dash; AIShield (https://boschaishield.com/)

National Institute of Standards and Technology

Docket Number: 231218-0309

About AIShield:

AIShield, a Bosch startup recognized by Gartner, stands at the forefront of cybersecurity for AI systems. AIShield's AI security technology is backed by over 45 patents and has been used globally by over 40 organizations in the automotive, manufacturing, banking, telecommunications, and healthcare industries since 2022. AIShield helps organizations to mitigate AI security risks before and after deployment and make AI systems (including Generative AI) resilient, and secure. This helps to improve the safety and trustworthiness of AI systems, as well as compliance with AI regulations and cybersecurity guidelines. The vision for the company is to empower organizations across industries to adopt AI with confidence, securing over 1000 AI systems globally by 2025 and beyond.

AIShield has actively participated in the development of guidelines, standards, best practices, and benchmarks for assessing the security and safety of AI systems in India and globally. This has been achieved by contributing to working groups of MITRE ATLAS, NASSCOM, DSCI (Data Security Council of India), DoT - Government of India, BIS (Bureau of Indian Standards), ETSI (European Telecommunications Standards Institute), FDA among others. AIShield has garnered accolades such as the CES Innovation Awards for 2023 & 2024 and the IoT World Congress Award, Black Hat MEA Excellence Award and has been featured in Gartner's AI TRiSM Market Guide, and AWS Generative AI Center of Excellence. Recognized by the OECD and cited in the G7 Hiroshima declaration, our solutions help organizations implement principles of international standards on cybersecurity of AI systems.

Context

The National Institute of Standards and Technology (NIST) is seeking information to assist in carrying out several of its responsibilities under the Executive order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence issued on October 30, 2023. Among other things, the E.O. directs NIST to undertake an initiative for evaluating and auditing capabilities relating to Artificial Intelligence (AI) technologies and to develop a variety of guidelines, including for conducting AI red-teaming tests to enable deployment of safe, secure, and trustworthy systems.

AIShield seeks to make valuable contributions to this endeavor.

Table of Contents

Scope of AIShield Information in this Document with respect to RFI

- #1 Developing Guidelines, Standards, and Best Practices for AI Safety and Security.
    - Developing a companion resource to the AI Risk Management Framework (AI RMF)
        - In terms of assessment, evaluation, monitoring risks.
    - Creating guidance and benchmarks for evaluating and auditing AI capabilities, with a focus on capabilities and limitations through which AI could be used to cause harm.
        - tools for evaluating the capabilities, limitations, and safety of AI technologies.
- #3 Advance Responsible Global Technical Standards for AI Development
    - Strategies for driving adoption and implementation of AI-related international standards.

Developing Guidelines, Standards, and Best Practices for AI Safety and Security

Developing a companion resource to the AI Risk Management Framework (AI RMF) in terms of assessment, evaluation, monitoring risks

Background discussions on risks and approach to evaluation of safety

What risks are anticipated due to the current and future use of AI?

The integration of AI in telecommunications, while promising, presents challenges, particularly in security.

- **Data Privacy Concerns**: Risks of data breaches with AI processing large volumes of user data.
- **AI Model Attacks**: Vulnerability of AI models to tampering, leading to erroneous decisions.
  - Data Poisoning Attacks: Manipulating training data to alter model behavior.
  - Input Manipulation Attacks: Deceiving models with crafted input data.
  - Membership Inference Attacks: Determining if data was in the training set.
  - Model Inversion/Data Reconstruction: Estimating training data from model interactions.
  - Model Supply Chain Attacks: Compromising models during their lifecycle.
- **AI Supply Chain Complexity**: Increased complexity in managing AI supply chain, vendor selection, and model assessment.
- **Reliance on External AI Solutions**: Dependency on third-party AI solutions or open-source AI models, raising security concerns.
- **AI-Driven Cyber Threats**: Emergence of sophisticated AI-powered malware and cyberattacks.
- **Regulatory Challenges**: Compliance with data privacy and regulatory obligations in AI adoption.

For detailed insights into these risks, please refer to the Appendix, Section I.

When evaluating the robustness of AI systems, what could be a potential approach?

When considering the robustness of AI systems, given the significant impact and critical nature of risks associated with AI systems, a security-centered responsible AI approach is pivotal. This approach leverages system insights (vulnerability, boundary conditions and loopholes) to effectively address key aspects such as fairness, explainability, reliability, scalability, and trust.

We propose a two-step evaluation process:

1. **Pre-Deployment Evaluation**: This phase involves a comprehensive assessment of a deployment-ready AI system before its actual deployment. It serves as a point-in-time analysis to evaluate the system against the dimensions mentioned above. During this phase, the model's performance limits, along with its operational boundaries and thresholds, are thoroughly examined. Meeting the success criteria at this stage is crucial for the AI system to be deemed production-ready.

2. **Real-Time Assessment in Production**: Once the AI system has been vetted, it undergoes continuous real-time monitoring during its operational phase. This includes vigilant tracking

of both incoming data and any variations in the AI model's performance relative to the established thresholds. This ongoing monitoring facilitates the activation of system-triggered alerts, notifications, and fallback strategies, which are essential for enhancing the system's overall trustworthiness.

How can one assess whether an AI system has a sufficient fallback plan for adversarial attacks or unexpected situations? How can real-world challenges be simulated during testing?

Assessment against AI System vulnerabilities against adversarial attacks and timely defense / mitigation of risks due to these attacks are essential. A good evaluation of the fallback / mitigation strategy and if missing, an effective implementation of such a strategy must be strongly considered during model deployment.

A risk-based approach, where the potential risks are identified, and the AI System is thoroughly evaluated against the identified risks. This should also ensure the risk-identification of the AI System together with its interfaces, interactions, and the deployment environment. Using the two-step evaluation approach (explained in B4) in the context of AI Security Testing:

- Step-1: Pre-Deployment Evaluation: Consideration of Real-world setting during testing.
    - Black-Box / Gray-Box Attacks: To ensure fair and robust evaluation of the AI Systems against adversarial attacks, a realistic evaluation representing real-world scenarios are essential. In a black-box attack scenario, the attacker has no knowledge about the system or the data. In a gray-box attack scenario, the attack gains information of the AI System and/or about its environment through Reconnaissance.
    - Attacker Proficiency: It is necessary to consider the highest proficiency of the attacker in the test phase. The attacker could use the AI attack techniques from the literature, open-source or use advanced AI based techniques. The test simulation setup should consider these upfront and test the AI System against these against and potential advanced attacks. For example: The vulnerability assessment platform of AIShield uses attack strategies from literature, open-source, proprietary attack techniques and Generative AI based attack approach, thereby ensuring state-of-the-art test coverage.
- Step-2: Real-time assessment of the AI System in production:
    - Step-1 exposes the vulnerabilities / loopholes of the AI System and provides insights on the conditions where the AI System could potentially fail. These vulnerabilities pose a real threat to the AI System in production. The fallback strategies must be defined once during the model development phase and once after the model evaluation phase. An efficient and timely implementation of the strategies in this step (Step-2), together with timely notification of the system failure is important.

Companion resource to the AI Risk Management Framework (AI RMF)

Considering the NIST AI RMF Playbook as reference, AIShield provides actionable steps and toolkit for realization of the framework. The resources and tools enable organizations to implement a “Govern” – centric risk management approach. AIShield emphasizes the methods of Risk management including Risk Assessment, Risk Evaluation and Mitigation to the “Map” – “Measure” – “Manage” phases in the RMF framework.

The tools *(Ref: Figure2)* align to the AI development lifecycle for both Predictive AI (ML and DL Models) and Generative AI (such as LLMs). The tools establish a methodology whereby the risk management team provide the multiple artefacts (notebooks, Models) as input. The outputs consist of results from thorough evaluation of the provided components and defense modules. The evaluation results enable as evidence for risk assessment. The defense modules can be implemented in the deployment / monitoring step. The mitigation measures bring down the overall risk, allowing an acceptable residual risk for organizations.

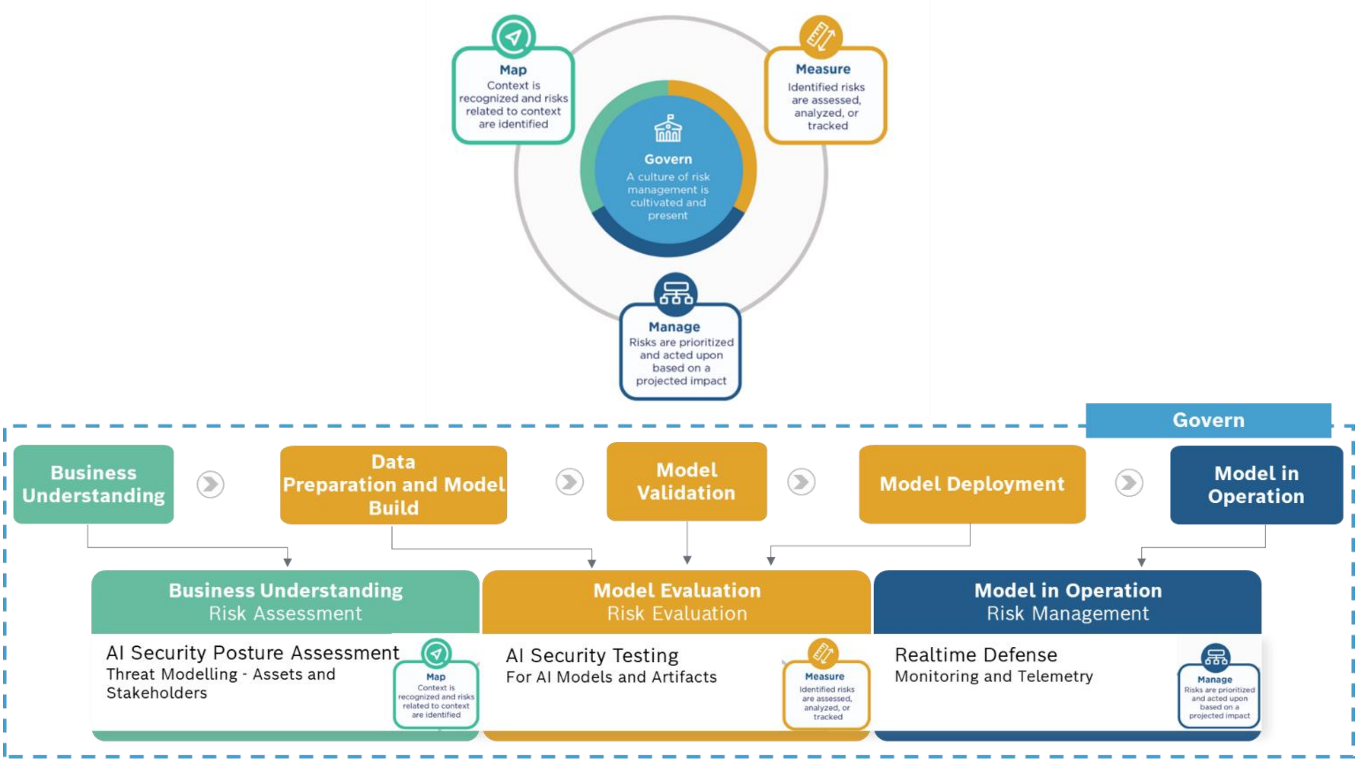

Map
Context is recognized and risks related to context are identified
Measure
Identified risks are assessed, analyzed, or tracked
Govern
A culture of risk management is cultivated and present
Manage
Risks are prioritized and acted upon based on a projected impact
Govern
Business Understanding
Data Preparation and Model Build
Model Validation
Model Deployment
Model in Operation
Business Understanding
Risk Assessment
AI Security Posture Assessment
Threat Modelling - Assets and Stakeholders
Map
Context is recognized and risks related to context are identified
Model Evaluation
Risk Evaluation
AI Security Testing
For AI Models and Artifacts
Measure
Identified risks are assessed, analyzed, or tracked
Model in Operation
Risk Management
Realtime Defense
Monitoring and Telemetry
Manage
Risks are prioritized and acted upon based on a projected impact

Figure 1: Mapping NIST AI RMF Playbook Principles to AI Development Workflow

Background discussions on AI System Robustness: Classification, Steps, Metrices and Roles

How can AI systems be categorised to facilitate their safety assessment?

In our view, AI systems must be categorized based on their characteristics, interaction, and deployments to have a holistic understanding on the assessment. Based on the above factors, we categorize AI Systems into – Machine Learning, Deep Learning and Generative AI Systems. Below comparison highlights their properties relevant for robustness assessment:

	Machine Learning (ML – AI 1.0)	Deep Learning (DL – AI 2.0)	Generative AI (GenAI – AI 3.0)	Relevance to Assessment

Model Complexity	Makes simple Linear Relations	Makes Neural Network based, Deep, Non-Linear and Complex Relations	Makes Very deep and Complex relations	Simpler models pose the properties of Explainability, transparency and accountability. While complex models such as DL and GenAI are usually considered blackbox.
Model Training Phase	Requires human intervention for hand-picking the features	Mistake and self-correction Method	Human in the Loop Feedback System	Having human intervention before or during the model training ensures certain levels of safety, fairness, reliability, and accountability of the System.
Dataset	Typically, finite number of hand-labelled and attributable samples		Potentially (infinite) large amount of non-attributable data	Attributable limited data used to train the AI system provides the possibility to assess the dataset from the perspective of – Security (Data Poisoning), Safety (Data boundaries), Resilience.
Deployment	Systems with or without GPU capabilities		Typically, Need specialized GPU	Scalability of the system has a direct correlation to the required deployment infrastructure.
Interactions	Typically, interact with other machines are part of decision making or automation. They also aid humans by making the decision-making process easier		Users interact directly with the GenAI Systems.	Both the interaction – with Humans or other machines – are going to have direct and indirect impacts leading to Security, Safety, privacy, fairness risks.

What essential steps or phases should be standardized for assessing AI system robustness?

Establishing a structured approach covering the essential steps or phases is crucial to standardize the assessment process. The traditional system assessment approaches (such as system definition, requirement analysis, data management, benchmarking, testing and maintenance) become anyway important. In addition, it is important to consider the steps and phases that become relevant due to the inclusion of AI in the System.

Below points emphasize on how the traditional steps and phases can now be used for AI Systems.

- Business Understanding: Together with the business objectives, goals, and stakeholder identification, it is essential to identify the risks and set clear objectives for AI Robustness.
- Data Management: In the data management phase, special emphasis must be given to assessment of data for security, safety, fairness, privacy, and accountability.
- Model and Data Evaluation: For the identified risks and objectives for AI Robustness, it is necessary to evaluate both the Data and the Model through appropriate metrics. The evaluation results must be audited and approved by identified stakeholders.
- Continuous monitoring and AI Robustness Management: Monitor system performance and security continuously in deployment, regularly update the system with necessary mitigation strategy to adapt to newer challenges.

What key metrics or performance indicators should be considered when evaluating AI system robustness?

The AI System evaluation metrics and performance indicators are subjective to the specific AI System. It is important to consider the AI System's functions, its impact due to compromised robustness on individuals, society and on organizations. Different evaluation metrics become important starting from system performance to resistance against security attacks.

Functional Performance	Accuracy, Precision, Errors (Root Mean Squared Error, Mean Average Error), F1 Score – to ensure reliable functional performance
Timing Performance	Latency and Response time – essential for time-critical systems like in Telecom
Security	Model Extraction Attack: Relative Model Accuracy (Accuracy of the extracted model against the original model) Model Evasion Attack: Attack Efficacy (% of Data samples misclassified due to the attack) Poisoning Attack: Poisoning Percentage (% of Data potentially poisoned by the attacker) Sponge Attack: QoS at constrained latency, increased Latency due to attack samples
Drift	Model Drift: Metric of the amount of drift against the preset (expected) outcome. Data Drift: Metric of the amount of drift noticed in the incoming data during production.
Fairness and Bias	Sensitive Class True Positive Rate, Demographic Parity Check

What roles are envisioned for government, standards organizations, and regulators in ensuring AI robustness?

Artificial intelligence (AI) is rapidly transforming the telecom and digital infrastructure landscape, offering numerous benefits but also introducing potential risks. To ensure the robustness and trustworthiness of AI in these domains, governments, standards organizations, and regulators play crucial roles:

Government

- Develops policy frameworks and regulations that promote responsible AI development and deployment.
- Provides funding and resources for AI research, focusing on enhancing robustness in telecom applications.

Standards Organizations

- Establishes industry standards for AI robustness, ensuring consistent and reliable AI performance.
- Facilitates the adoption of these standards and provides guidance on best practices for robust AI implementation.

Regulators

- Enforces compliance with AI robustness standards to protect the safety, privacy, and reliability of telecom networks and digital infrastructure.
- Monitors and evaluates the implementation of AI technologies in telecom to identify potential risks and ensure responsible AI usage.

By working together, these stakeholders can foster a robust and trustworthy AI ecosystem in the telecom and digital infrastructure domains, maximizing the benefits of AI while mitigating potential risks. For a more detailed explanation, please refer to Appendix, Section II.

Tools for evaluating the capabilities, limitations, and safety of AI technologies.

AIShield provides a comprehensive suite of tools and capabilities for AI Vulnerability assessment and mitigation across the AI Lifecycle – including the development and the deployment phases. The different toolkits cover the need for assessment across model development artefacts – for supply chain vulnerability assessment, AI models and Large Language Models (LLMs).

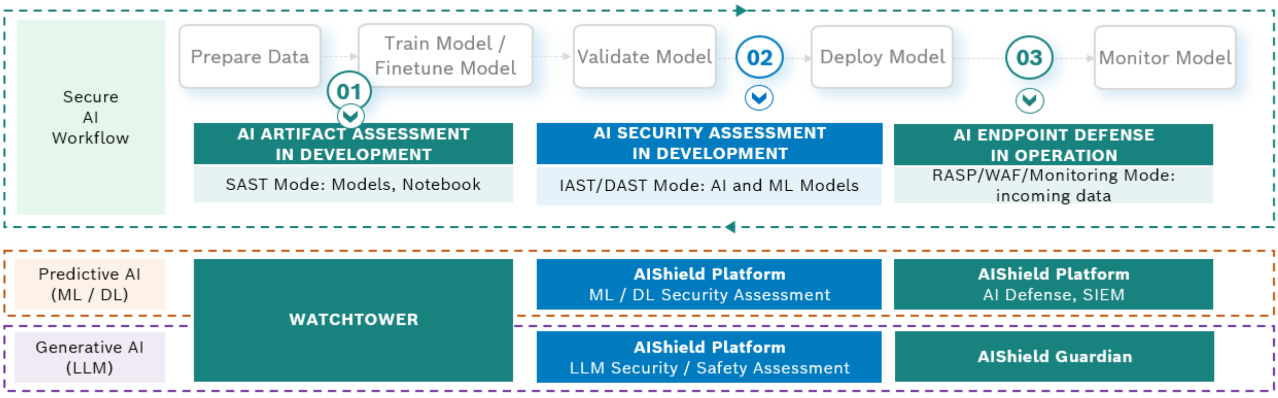

Secure
AI
Workflow
Prepare Data
Train Model / Finetune Model
Validate Model
Deploy Model
Monitor Model
01
02
03
AI ARTIFACT ASSESSMENT IN DEVELOPMENT
SAST Mode: Models, Notebook
AI SECURITY ASSESSMENT IN DEVELOPMENT
IAST/DAST Mode: AI and ML Models
AI ENDPOINT DEFENSE IN OPERATION
RASP/WAF/Monitoring Mode: incoming data
Predictive AI (ML / DL)
Generative AI (LLM)
WATCHTOWER
AIShield Platform
ML / DL Security Assessment
AIShield Platform
AI Defense, SIEM
AIShield Platform
LLM Security / Safety Assessment
AIShield Guardian

Figure 2: Mapping of AIShield tools to AI Development Workflow

**SAST – Static Application Security Testing; DAST – Dynamic Application Security Testing; IAST – Interactive Application Security Testing; RASP – Real time Application Security Protection; WAF – Web Application firewall

The table below shows the different tools mapped across the AI Development workflow and provides useful links to explore them.

Tool	Description	Link
WATCHTOWER	Vulnerability scan for AI Models and Notebooks.	https://github.com/bosch-aisecurity-aishield/watchtower
AIShield Platform	AI Model Vulnerability assessment and defense Model generation	https://www.boschaishield.com/product/
AIShield Guardian	Guardrails for safe and secure deployment of LLMs in Enterprises	https://boschaishield.co/guardian, https://oecd.ai/en/catalogue/tools/aishield-guardian

Advance Responsible Global Technical Standards for AI Development

Strategies for driving adoption and implementation of AI-related international standards.

Contours of a possible standard for assessing AI robustness

What key features should a standard for assessing AI systems for robustness include?

In an era where AI systems are integral in most industries, establishing a robust standard for assessing these systems is paramount. This standard must address a spectrum of considerations from risk identification to stakeholder involvement, ensuring AI systems are not only technologically advanced but also safe, reliable, and ethically responsible.

Key Features for AI Robustness Standard:

- **Risk-Based Approach**: Comprehensive assessment of potential risks like data privacy, security breaches, and decision-making biases.
- **Clear Risk Assessment Methodologies**: Detailed strategies for evaluating and mitigating risks.
- **Mitigation Measures**: Special focus on robust solutions for high-risk scenarios.
- **Thorough Documentation**: Essential for auditing and compliance of AI systems.
- **Advanced Tools for Risk Management**: Incorporating AI analytics and simulation tools for proactive risk management.
- **Continuous Monitoring and Updates**: Ensuring AI systems stay current with evolving threats and technologies.
- **Stakeholder Involvement**: Collaboration among telecom providers, AI developers, regulatory bodies, and end-users.
- **Human Oversight Interfaces**: Adaptable oversight mechanisms for different AI applications.
- **Global Standardization Efforts**: Engagement in international AI standardization and ethical use.
- **Incident Response Plans**: Protocols for quick recovery from AI system failures or breaches.

For detailed insights and further elaboration, please refer to the Appendix, Section III. Additionally, we have included a self-assessment checklist in Appendix, Section III that can be used for ensuring AI systems robustness across each phase of the AI lifecycle.

What considerations should be taken into account to ensure that the standard for assessing AI robustness remains adaptable to future advancements in AI?

To ensure the standard for assessing AI robustness remains adaptable to future advancements in AI and telecom technologies, the following considerations are crucial:

- Establish a dynamic framework for generative AI risks.
- Introduce strict controls and human oversight for AI systems.
- Create strategies to protect against AI model attacks.
- Foster international cooperation and encourage industry contributions.
- Actively incorporate emerging AI technologies and standards.
- Design scalable and adaptable standards for future technologies.

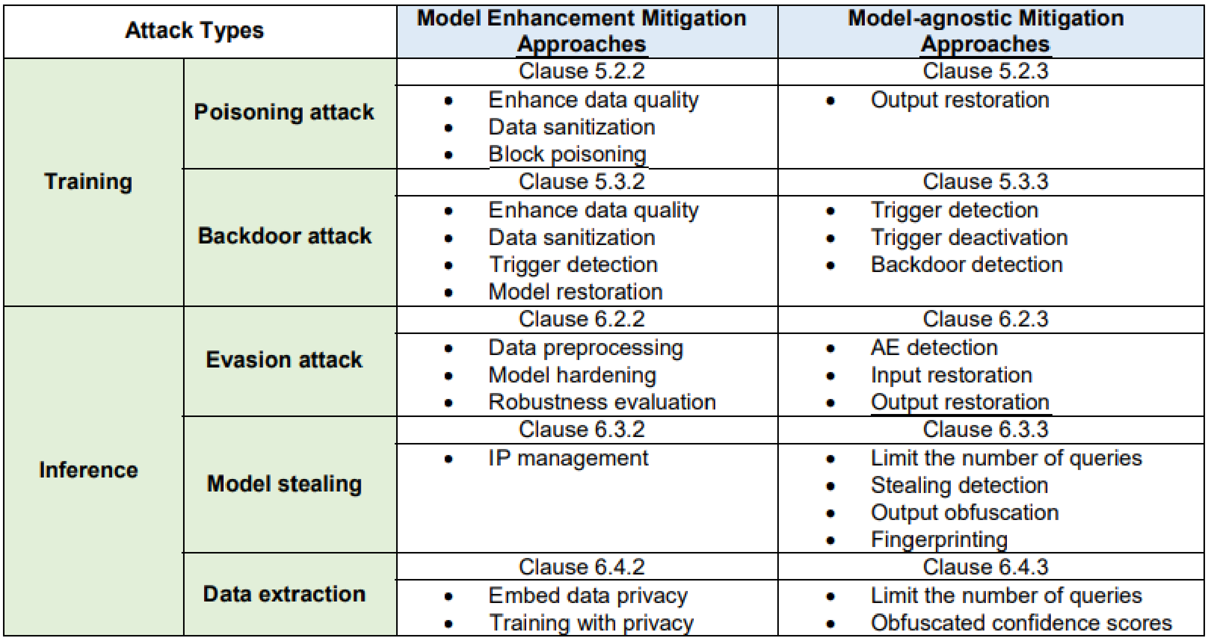
Appendix

I. Risks are anticipated due to the current and future use of AI

The integration of AI in telecommunications, while promising, presents challenges, particularly in security.

1. **Data Privacy**: With AI systems processing vast amounts of user data, data breaches are an inherent risk. To protect user privacy, telecom providers must implement robust encryption and adhere to data protection regulations. The AI pipelines is an additional attack surface. An important risk factor in the additional attack surface is the presence of production data in the engineering process. To train and test a working model, data scientists need access to real data, which may be sensitive. This is different from non-AI engineering in which typically the test data can be either synthesized or anonymized.
2. **AI Model Attacks**: As AI models become central to telecom operations, there is a risk of tampering by adversaries. Such attacks can cause the AI system to make erroneous decisions, possibly disrupting telecommunications services.
    a. **Data poisoning attack**: by changing training data (or labels of the data), the behavior of the model can be manipulated. This can either sabotage the model or have it make decisions in favor of the attacker.
    b. **Input manipulation attack**: fooling models with deceptive input data. This attack can be done in three ways: 1) by experimenting with the model input (black box), 2) by introducing maliciously designed input based on analysis of the model parameters (white box), and 3) by basing the input on data poisoning that took place.
    c. **Membership inference attack:** given a data record (e.g. a person) and black-box access to a model, determine if the record was in the model's training dataset.
    d. **Model inversion attack**, or *data reconstruction*: by interacting with or by analyzing a model, it can be possible to estimate the training data with varying degrees of accuracy.
    e. **Model supply chain attack**: attacking a model by manipulating the lifecycle process to actual use. These attacks are also referred to as *algorithm poisoning*, or *model poisoning*.
3. **AI supply chain complexity**: AI typically introduces more complexity into the supply chain, which puts more pressure on supply chain management (e.g. vendor selection, pedigree and provenance, third-party auditing, model assessment, patching and updating). The problem is increased by the threat of the various model attacks, in combination with the fact that model behavior can typically not be assessed through static analysis.
4. **Reliance on External AI Solutions**: Numerous telecom operators rely on AI solutions from third parties. It is crucial to ensure the security of these external systems, as vulnerabilities in these systems can compromise the telecom operator's overall security.
5. **AI-Driven Threats**: With the rise of AI, cyber threats driven by AI have also increased. These include malware powered by artificial intelligence, automated cyberattacks, and sophisticated phishing campaigns. Telecom operators must be prepared to counter these threats of the next generation.
6. **Regulatory**: Telecom industry's adoption of AI/ML technologies is constrained by data accessibility, privacy, and regulatory obligations.

Table 1: Mitigation Strategy Framework



Attack Types		Model Enhancement Mitigation Approaches	Model-agnostic Mitigation Approaches
Training	Poisoning attack	Clause 5.2.2 • Enhance data quality • Data sanitization • Block poisoning	Clause 5.2.3 • Output restoration
	Backdoor attack	Clause 5.3.2 • Enhance data quality • Data sanitization • Trigger detection • Model restoration	Clause 5.3.3 • Trigger detection • Trigger deactivation • Backdoor detection
Inference	Evasion attack	Clause 6.2.2 • Data preprocessing • Model hardening • Robustness evaluation	Clause 6.2.3 • AE detection • Input restoration • Output restoration
	Model stealing	Clause 6.3.2 • IP management	Clause 6.3.3 • Limit the number of queries • Stealing detection • Output obfuscation • Fingerprinting
	Data extraction	Clause 6.4.2 • Embed data privacy • Training with privacy	Clause 6.4.3 • Limit the number of queries • Obfuscated confidence scores

Attacks Types: Risks to AI Systems | Reference: ETSI Securing Artificial Intelligence (SAI); Mitigation Strategy Report
[*https://www.etsi.org/deliver/etsi_gr/SAI/001_099/005/01.01.01_60/gr_SAI005v010101p.pdf*](https://www.etsi.org/deliver/etsi_gr/SAI/001_099/005/01.01.01_60/gr_SAI005v010101p.pdf)

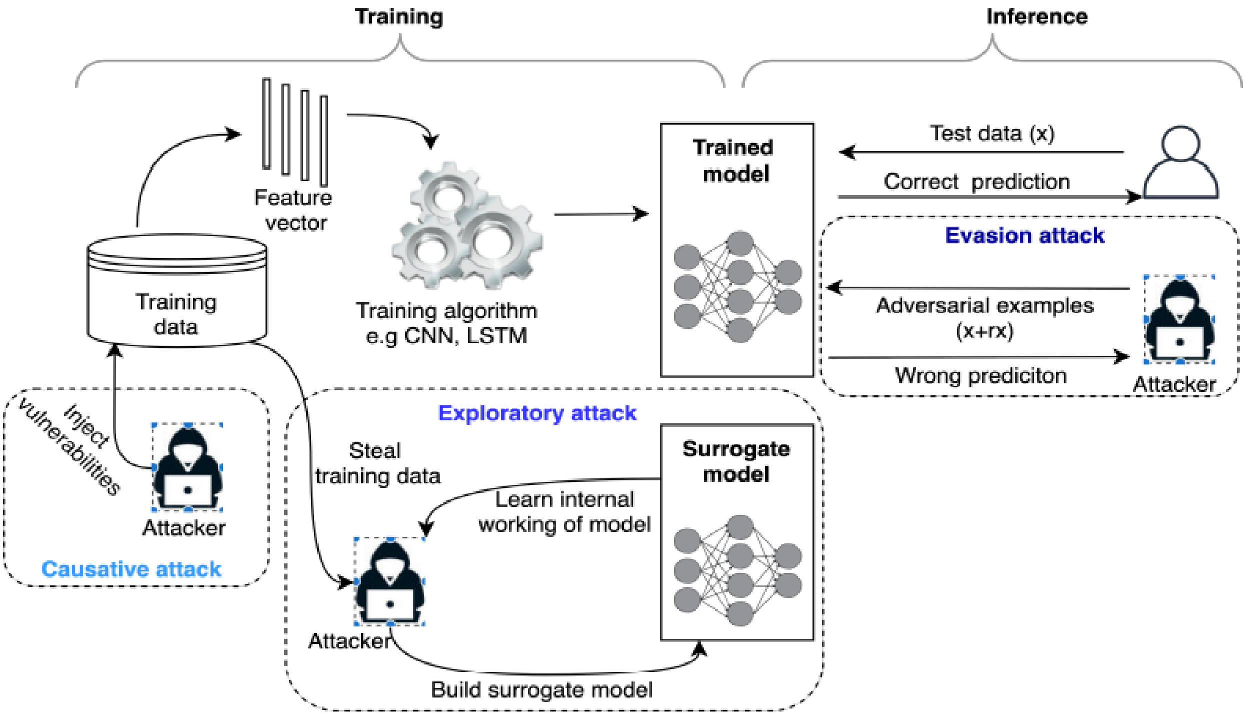
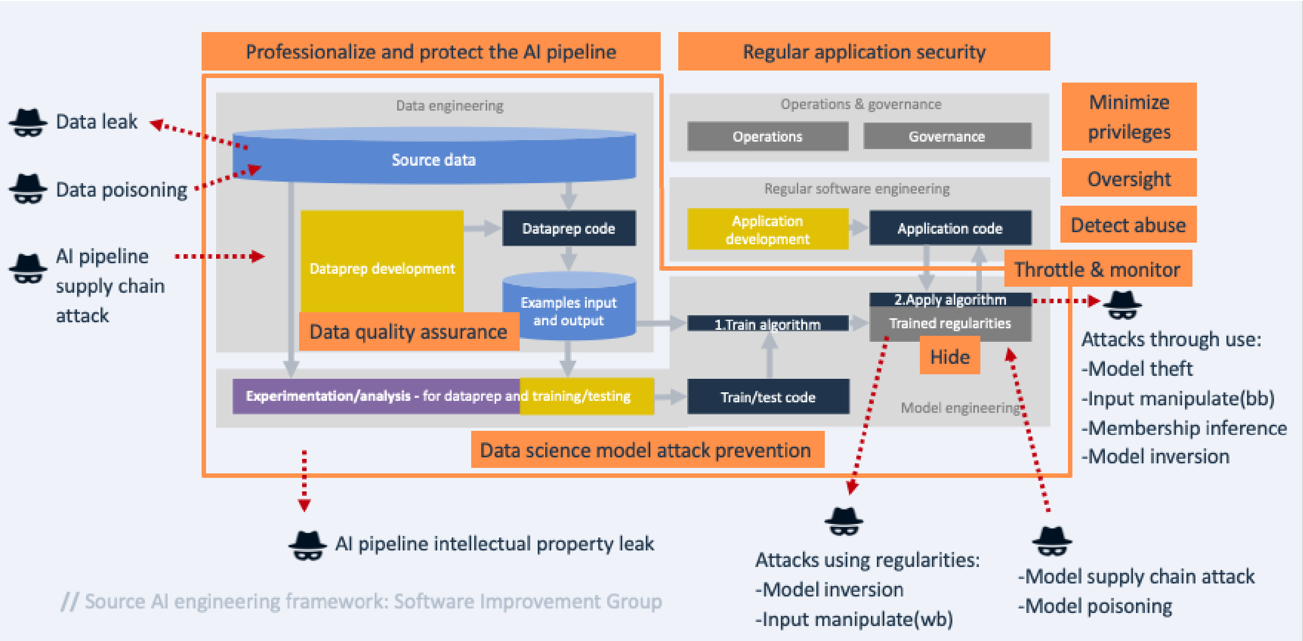


Fig. 3. Attacks in adversarial machine learning [136].

Attacks on Adversarial Machine Learning | Adversarial Machine Learning in Wireless Communications using RF Data: A Review | Reference: https://arxiv.org/pdf/2012.14392.pdf

Generic AI/ML Risks - Top 10 Machine Learning Security Risks (OWASP)



AI security risks are visualized in the diagram, together with key mitigation (orange) | Reference: https://owasp.org/www-project-ai-security-and-privacy-guide/

II. Roles envisioned for government, standards organisation, and regulators in ensuring AI robustness in telecom networks and digital infrastructure.

As artificial intelligence (AI) continues to permeate various aspects of our lives, including telecommunications networks and digital infrastructure, it is crucial to ensure the robustness and trustworthiness of these systems. AI-powered technologies offer numerous benefits, but they also introduce potential risks, such as vulnerabilities to cyberattacks, biases, and unintentional harm. Therefore, it is imperative for governments, standards organizations, and regulators to play a proactive role in ensuring the responsible and ethical development and deployment of AI in the telecom and digital domains.

Government

Governments have a critical responsibility in establishing and enforcing policies and regulations that promote AI robustness in telecom networks and digital infrastructure. This includes:

- **Developing clear guidelines and frameworks for AI development and deployment**: Governments can provide a roadmap for AI development, outlining principles and practices that foster responsible and ethical AI development.

- **Establishing oversight mechanisms for AI systems**: Governments can establish regulatory bodies or agencies tasked with monitoring and evaluating AI systems, ensuring they adhere to established standards and guidelines.

Promoting research and innovation in AI robustness: Governments can invest in research and development initiatives focused on enhancing AI robustness, supporting the development of new techniques and tools for detecting and mitigating AI vulnerabilities.

Standards Organizations

Standards organizations play a crucial role in defining technical specifications and best practices for AI systems. In the context of telecom networks and digital infrastructure, standards organizations can contribute to AI robustness by:

- **Developing standards for AI development and testing**: Standards organizations can establish standardized methodologies for developing and testing AI systems, ensuring they meet certain levels of robustness and reliability.
- **Promoting the use of open-source AI tools**: Standards organizations can encourage the adoption of open-source AI tools and frameworks, facilitating transparency and collaboration in AI development and enhancing the ability to identify and address potential vulnerabilities.
- **Establishing standards for AI data governance**: Standards organizations can define guidelines for managing and protecting AI data, ensuring data privacy, security, and responsible data usage.

Regulators

Regulators oversee the operation of telecom networks and digital infrastructure, and they play a vital role in ensuring the safety, reliability, and security of these systems. In the context of AI, regulators can contribute to AI robustness by:

- **Enforcing regulations related to AI security and robustness**: Regulators can mandate that AI systems used in telecom and digital infrastructure are secure and robust, allowing for better understanding of their decision-making processes.
- **Establishing requirements for AI auditing and monitoring**: Regulators can require regular auditing and monitoring of AI systems deployed in telecom networks and digital infrastructure, identifying, and addressing potential risks and vulnerabilities.
- **Enacting measures to prevent the misuse of AI**: Regulators can establish safeguards to prevent the misuse of AI in telecom networks and digital infrastructure, such as prohibiting the use of AI for discriminatory practices or unauthorized data collection.

By working together, governments, standards organizations, and regulators can create an ecosystem that fosters the development and deployment of robust and trustworthy AI in telecom networks and digital infrastructure, ensuring that AI technologies are used responsibly and ethically for the benefit of society.

III. Key features of a standard for assessing AI systems for robustness, specifically focused on telecom networks and other digital infrastructure

In an era where AI systems are integral to telecom networks and digital infrastructure, establishing a robust standard for assessing these systems is paramount. This standard must address a spectrum of considerations from risk identification to stakeholder involvement, ensuring AI systems are not only technologically advanced but also safe, reliable, and ethically responsible.

The following are the key features the standard should possess:

1. **Risk-Based Approach and Potential Risks**: The standard should include a comprehensive overview of potential risks associated with AI systems in telecom and digital infrastructure. This includes risks related to data privacy, security breaches, system failures, biases in decision-making processes, and any other risks that could affect the reliability and safety of telecom networks. This includes:
    a. Enumerating high-risk applications.
    b. Setting clear requirements for AI systems, especially in high-risk domains.
    c. Defining obligations for AI users and providers in these areas.
    d. Proposing a conformity assessment prior to AI deployment.
    e. Implementing enforcement and governance structures post-deployment.
    f. Mandating third-party auditing and certification.
    g. Introducing measures to prevent and track AI model leaks.
    h. Expanding funding for technical AI safety research.
2. **Clear Methodologies to Assess Risks**: The standard should provide clear, detailed methodologies for assessing the identified risks. This would involve guidelines on how to evaluate the severity and likelihood of each risk, along with strategies to mitigate them. These methodologies should be adaptable to various types of AI applications within the telecom sector. there is an urgent need to adopt a regulatory framework by the Government that should be applicable across sectors. The regulatory framework should ensure that specific AI use cases are regulated on a risk-based framework where high risk use cases that directly impact humans are regulated through legally binding obligations.
3. **Mitigation Measures:** The standard should outline specific mitigation measures for identified risks, emphasizing the need for robust solutions in high-risk scenarios.
4. **Documentation of Evidence for Auditing AI Systems**: There should be a requirement for thorough documentation of AI systems, including their design, development, deployment, and maintenance processes. This documentation would be crucial for auditing purposes, ensuring that the AI systems comply with the set standards and allowing for traceability in case of any issues.
5. **Supporting Tool-Assisted Risk Management**: The standard should also encourage or mandate the use of advanced tools for risk management. These could include AI-powered analytics tools for continuous monitoring of AI systems, simulation tools for stress-testing AI applications under various scenarios, and other technological solutions that assist in proactively managing and mitigating risks.
6. **Continuous Monitoring and Updating**: Implement continuous monitoring mechanisms for AI systems and mandate regular updates to ensure they adapt to new threats, technologies, and evolving industry standards.
7. **Stakeholder Involvement**: Encourage active involvement from all stakeholders, including telecom providers, AI developers, regulatory bodies, and end-users, in developing and maintaining these standards. The AI ecosystem has multiple stakeholders- private sector, research, government, legal bodies, regulators, standard setting bodies, etc. The regulatory principles are expected to serve these stakeholders of the AI ecosystem. The AI technology is not confined to a sector. Moreover, the issues involved are wide and complex. It would require consultation with various stakeholders on various aspects of AI. Therefore, there should be a mechanism for an elaborate consultation with all the concerned stakeholders while formulating or updating AI regulations and guidelines.

8. **Interface for Human Agency and Oversight**: Depending on the use case, different interfaces for human oversight may be necessary, such as GUI, CLI, or API. For instance, an AI system predicting network congestion might use a graphical interface integrated into a regular network dashboard, including alerts for deteriorating AI performance and actionable information for NOC engineers.
9. **Standardization and Engagement in International AI Standards**: Engaging in the development of international AI standards is crucial. The U.S. Department of State's focus on creating standards for AI technologies through international partnerships and the Global Partnership on AI (GPAI) is a key example. This global approach ensures consistency in standards and helps in implementing trustworthy AI technologies.
10. **Incident Response and Recovery Plans**: Include protocols for responding to AI system failures or breaches, outlining steps for quick recovery, and minimizing disruption in telecom services.

Self-Assessment Checklist

A proposed self-assessment checklist that may be included in the standard for safety and robustness of AI systems:

PHASE	CHECKLIST ITEM
PLANNING & DESIGN	
	Have you analyzed the risk factors that may occur during the lifecycle of the AI system?
	Did you take measures to eliminate, prevent, or minimize the effects of the risk factors?
DATA COLLECTION & PROCESSING	
	Removal of abnormal data to ensure data robustness • Have you identified data outliers and checked for normality and errors?
	Have you made an effort to protect the data against attacks? • Did you provide defense measures against attacks such as data poisoning and evasion?
AI MODEL DEVELOPMENT	
	Ensuring security and compatibility of open-source library • Have you verified the security and compatibility of the open-source library? ○ Did you confirm the license, security vulnerability, and compatibility of the open-source library being used?
	Establishment of response countermeasures against AI model attacks • Did you introduce response measures against model extraction attacks? ○ Did you apply a defense technique to prepare for model extraction attacks?
	Implementation of safe mode of AI system Did you apply safe mode in the case of an attack, a deterioration in the performance, or social issue? • Do you have a policy to deal with exceptions to a problematic situation? • Have you applied a security mechanism to strengthen the security of the AI system? • When a problem arises, do you consider human intervention? • Do you provide guidance and responses regarding expected user errors?

	Did you generate a report when a problem occurred with the AI system? • Have you established a reporting procedure for ethical issues, such as prejudice or discrimination? Have you set up indicators and procedures to assess the performance degradation of the system?
OPERATION & MONITORING	
	Securing traceability of AI system • Have you established measures to track and respond to the decision-making of the AI system? • Do you regularly manage the records of changes to the training data? • Do you periodically update the history of the training data being managed? ○ When securing new data, do you reconduct a performance and security evaluation of the AI model?