IS2109 Information Systems Security

AI RISK MANAGEMENT FRAMEWORK

National Institute of Standards and Technology

Prepared By

W.C.M. Gunathilaka – 21020312

D.G.S. Hemathilake – 21020361

S.M.W.P.W.Bandara – 21020116

R.M.T.P.Rathnayaka – 21020795

Table of Contents

Introduction to NIST

The National Institute of Standards and Technology (NIST) is a non-regulatory federal agency within the U.S. Department of Commerce. Established in 1901 as the National Bureau of Standards (NBS) and renamed in 1988, NIST's primary mission is to promote U.S. innovation and industrial competitiveness by advancing measurement science, standards, and technology in ways that enhance economic security and improve our quality of life.

NIST works to ensure that measurements, standards, and technologies are both reliable and accurate, which is crucial across a wide range of industries, including manufacturing, telecommunications, cybersecurity, healthcare, and environmental sciences. By providing the technical foundation for industries to innovate, NIST plays a pivotal role in facilitating product development, quality assurance, and commerce.

The core competencies of the National Institute of Standards and Technology (NIST) :

- Measurement Science: The practice of ensuring measurements are accurate, precise, and consistent, foundational for innovation and quality control.
- Rigorous Traceability: Ensuring all measurements can be reliably traced back to national or international standards, crucial for product safety and quality.
- Development and Use of Standards: Creating and promoting agreed-upon methods and criteria to ensure compatibility, safety, and performance across products and systems.

What we're gonna highlight here falls under Development and Use of Standards. Under **NIST's Standards Research & Development** they have,

- AI Risk Management Framework
- Cybersecurity Framework
- Post-Quantum Cryptography
- Advanced Communications

In this document we will dive deep into the AI Risk Management Framework and how it is applied in various organizations in their AI related development.

Overview of Artificial Intelligence (AI)

In computer science, artificial intelligence (AI) is the study of building intelligent machines that can carry out tasks that normally call for human intelligence. Creating machines with human-like perception, reasoning, learning, and behavior is the main objective of artificial intelligence (AI).

Key aspects of artificial intelligence:

- **Machine Learning:** Machine learning is a subset of AI that focuses on enabling machines to learn from data and improve their performance over time without being explicitly programmed. This involves techniques such as supervised learning, unsupervised learning, and reinforcement learning.
- **Natural Language Processing (NLP):** NLP enables computers to understand, interpret, and generate human language. It involves tasks such as language translation, sentiment analysis, and text summarization.
- **Computer Vision:** Computer vision enables machines to interpret and understand the visual world. It involves tasks such as object recognition, image classification, and facial recognition.
- **Expert Systems:** Expert systems are AI systems that emulate the decision-making ability of a human expert in a specific domain. They use a knowledge base and a set of rules to provide advice or make decisions.
- **Robotics:** Robotics combines AI with mechanical engineering to create robots capable of interacting with the physical world. AI enables robots to perceive their environment, make decisions, and adapt to changes in their surroundings.
- **Deep Learning:** Deep learning is a subset of machine learning that uses artificial neural networks with multiple layers to learn complex patterns in large amounts of data. It has achieved significant breakthroughs in areas such as image recognition, speech recognition, and natural language processing.
- **Ethical and Social Implications:** AI raises important ethical and societal questions, including concerns about job displacement, bias in algorithms, privacy issues, and the impact on warfare and security.

Risks in Artificial Intelligence(AI)

Artificial intelligence (AI) offers tremendous potential benefits, but it also presents several risks and challenges.

- **Job Displacement:** One of the most significant concerns is the potential for AI to automate tasks currently performed by humans, leading to job displacement and economic disruption. Industries such as manufacturing, transportation, and customer service are particularly vulnerable to automation.
- **Bias and Fairness:** AI systems can perpetuate or even amplify biases present in the data used to train them. This can lead to unfair outcomes in areas such as hiring, lending, and criminal justice. Ensuring fairness and mitigating bias in AI algorithms is a critical challenge.
- **Privacy and Security:** AI systems often rely on vast amounts of data, raising concerns about privacy and data security. Unauthorized access to sensitive data or breaches of AI systems could have serious consequences, including identity theft, surveillance, and manipulation.
- **Autonomous Weapons:** The development of autonomous weapons systems powered by AI raises ethical and legal concerns. There are fears that such weapons could lead to unintended escalation, civilian casualties, and violations of international humanitarian law.
- **Exacerbating Inequality:** The benefits of AI may not be equally distributed, leading to widening socioeconomic inequality. Access to AI technologies, education, and employment opportunities could further disadvantage marginalized communities.
- **Lack of Transparency and Accountability:** AI systems often operate as "black boxes," making it difficult to understand how they arrive at decisions. This lack of transparency can undermine trust and accountability, specially in critical applications such as healthcare and finance.
- **Existential Risks:** Some experts warn of the long-term existential risks posed by advanced AI systems surpassing human intelligence. These risks include scenarios such as AI systems with goals misaligned with human values or inadvertently causing harm due to unforeseen consequences.

Introduction to the NIST AI Risk Management Framework (RMF)

1. Framework Overview:

- Developed by the National Institute of Standards and Technology (NIST).
- It is intended for voluntary use and aims to enhance the ability to incorporate trustworthiness considerations into the design, development, use, and evaluation of AI products, services, and systems.
- Aims to manage risks associated with artificial intelligence (AI) in design, development, use, and evaluation of AI products, services, and systems.
- Released on January 26, 2023.
- Developed through a consensus-driven, open, transparent, and collaborative process.
- Its primary goal is to manage risks associated with artificial intelligence.

2.Functions of the Framework:

- Govern: Establish governance structures and processes.
- Map: Identify and assess AI-related risks.
- Measure: Quantify and evaluate risks.
- Manage: Implement risk mitigation strategies.

3.Companion Resources:

- NIST AI RMF Playbook: Practical guidance.
    - Provides practical guidance for implementing the framework.
    - Suggests ways to navigate and use the AI RMF.
    - Incorporates trustworthiness considerations into AI design, development, deployment, and use.
    - Link : NIST AI RMF Playbook | NIST

- AI RMF Roadmap: Path for AI risk management.
    - Outlines the strategic path for AI risk management.
    - Identifies key activities and milestones.
    - Link : Roadmap for the NIST Artificial Intelligence Risk Management Framework (AI RMF 1.0) | NIST

- AI RMF Crosswalk: Alignment with other AI resources and standards.
    - Provides alignment between the AI RMF and other relevant standards or frameworks.

    - Helps organizations understand how the AI RMF relates to existing practices.
    - Link : Crosswalks to the NIST Artificial Intelligence Risk Management Framework (AI RMF 1.0) | NIST

- Various Perspectives: Additional insights.
    - A collection of statements by companies, industry organizations, and advocacy groups.
    - Offers additional insights and viewpoints on the AI RMF.
    - Link : Perspectives about the NIST Artificial Intelligence Risk Management Framework | NIST

- AI RMF Explainer Video: Overview of the framework.
    - Provides an overview of the AI RMF.
    - Helps users understand its purpose and key components.
    - Link : Introduction to the NIST AI Risk Management Framework (AI RMF 1.0): An Explainer Video | NIST

4.Trustworthy and Responsible AI Resource Center:

- Launched on March 30, 2023.
- Facilitates implementation of the AI RMF and promotes international alignment.

5.Prior Documents:

- Second draft of the AI Risk Management Framework (August 18, 2022).
    - This draft represents an intermediate version of the framework.
    - It likely underwent revisions based on feedback and insights from stakeholders.
    - The purpose was to refine and enhance the framework's content before its final release.

- Initial draft of the AI Risk Management Framework (March 17, 2022).
    - The initial draft served as the foundational version of the AI RMF.
    - It laid out the fundamental concepts, functions, and principles.
    - Stakeholders likely provided input during the development process.

- Concept paper to guide development of the AI RMF (December 13, 2021).
    - The concept paper was a precursor to the formal framework.
    - It likely outlined the need for an AI-specific risk management approach.

    - It might have highlighted key challenges and considerations related to AI risk.

- Brief summary of responses to the July 29, 2021, Request for Information (RFI).
    - The RFI sought input from the public and experts.
    - Stakeholders provided feedback, insights, and recommendations.
    - The summary likely captured key themes, concerns, and suggestions related to AI risk management.

- Draft - Taxonomy of AI Risk (October 15, 2021).
    - The taxonomy aimed to categorize different types of AI-related risks.
    - It likely included risk dimensions such as bias, security, fairness, and explainability.
    - The purpose was to create a structured framework for understanding and addressing AI risks.

Components of the NIST AI RMF

1. Key Concepts

- Trustworthiness: Incorporating trustworthiness considerations into AI design and use.
    - Trustworthiness refers to the degree to which an AI system can be relied upon to perform its intended functions accurately, ethically, and safely.
      :
        - Accuracy: Ensuring that AI systems deliver reliable and dependable outcomes.
        - Safety: Prioritizing user safety and minimizing harm.
        - Ethics: Addressing ethical implications, fairness, and bias.
        - Transparency: Making AI processes transparent and explainable.
        - Security: Safeguarding against malicious attacks and vulnerabilities.

- Risk Management: Identifying, assessing, and mitigating AI-related risks.
    - Risk management involves systematically identifying, assessing, and mitigating risks associated with AI.
    - Process:
        - Identification: Recognize potential risks related to AI deployment.
        - Assessment: Evaluate the likelihood and impact of each risk.
        - Mitigation: Implement strategies to reduce or control risks.
        - Monitoring: Continuously monitor AI systems for emerging risks.
    - Balancing Act: Balancing innovation and benefits with risk mitigation.

- Voluntary Adoption: Organizations can choose to adopt the framework.
    - The AI RMF is not mandatory; organizations can choose to adopt it based on their needs and context.
    - Voluntary adoption allows flexibility in tailoring the framework to specific organizational requirements.
    - Organizations can align their risk management practices with the AI RMF voluntarily.

- Collaborative Process: Developed with input from various stakeholders.

- The AI RMF was developed through a collaborative effort involving various stakeholders:
  - Industry: Input from AI practitioners, researchers, and developers.
  - Government: NIST, regulatory bodies, and policymakers.
  - Academia: Researchers, educators, and experts.
  - Public: Public comments, workshops, and feedback.

- Collaboration ensures diverse perspectives and robust risk management practices.

2. Integration with NIST RMF

- The AI RMF aligns with the NIST Risk Management Framework (RMF).

  - The AI RMF builds upon the principles of the NIST RMF.
  - Here's how it adapts and enhances existing risk management practices for AI-specific contexts:
    - Tailored Controls: The AI RMF incorporates controls from NIST SP 800-53, which offers a comprehensive catalog of security controls. Organizations can select and implement controls specific to their AI systems' risk profiles and requirements.
    - Risk Assessment: The AI RMF aligns with the RMF's risk assessment process. However, it emphasizes AI-specific risk factors such as bias, explainability, and robustness.
    - Lifecycle Integration: The AI RMF integrates seamlessly into the system development life cycle (SDLC), ensuring that AI risk management is an integral part of AI system design, development, deployment, and maintenance.

- Adapts RMF principles to address AI-specific risks.

  - Alignment with NIST RMF:
    - The NIST RMF is a well-established framework for managing risks associated with information systems and organizations.
    - It provides guidelines for assessing, selecting, implementing, and monitoring security controls.
    - The AI RMF builds upon the NIST RMF, recognizing that AI introduces unique challenges and considerations.

    - AI-Specific Adaptations:The AI RMF tailors RMF principles to address the specific risks posed by artificial intelligence:
        - Contextualization: It considers the context in which AI systems operate, including data sources, user interactions, and AI-specific challenges.
        - Risk Assessment: The AI RMF evaluates AI-specific risks such as bias, fairness, explainability, and robustness.
        - Controls Selection: It incorporates controls from NIST SP 800-53 that are relevant to AI systems.
        - Lifecycle Integration: The AI RMF seamlessly integrates risk management practices into the AI system's lifecycle.

    - Enhancing Responsible AI:
        - By adapting RMF principles, the AI RMF ensures that AI risk management aligns with established best practices while addressing the unique challenges posed by artificial intelligence.
        - It promotes responsible AI development, deployment, and use by emphasizing trustworthiness, ethics, and societal impact.

- Enhances existing risk management practices.

    - By leveraging the NIST RMF principles, the AI RMF enhances AI risk management by:
        - Providing a structured approach to assess and manage AI risks.
        - Encouraging organizations to consider trustworthiness, ethics, and societal impact.
        - Facilitating collaboration between technical and non-technical stakeholders.
        - Ensuring that AI systems are secure, reliable, and privacy-preserving.

3. AI Risk Management Process

- Govern: Establish governance structures.

    - Purpose: The "Govern" phase focuses on establishing robust governance structures and processes for AI risk management.
    - Key Aspects:
        - Governance Structures: Define roles, responsibilities, and decision-making authorities related to AI risk.

        - Policies and Procedures: Develop clear policies, guidelines, and procedures for AI risk management.
        - Risk Oversight: Ensure that AI risks are monitored and addressed at all levels of the organization.
        - Communication Channels: Establish channels for communicating risk-related information.

- Map: Identify and assess AI-related risks.

    - Purpose: The "Map" phase involves identifying and assessing AI-related risks comprehensively.
    - Key Activities:
        - Risk Identification: Identify potential risks associated with AI systems, considering technical, ethical, and legal aspects.
        - Risk Assessment: Evaluate the likelihood and impact of identified risks.
        - Risk Context: Understand the context in which AI systems operate (e.g., data sources, user interactions).
        - Risk Documentation: Document risk profiles and maintain risk registers.

- Measure: Quantify and evaluate risks.

    - Purpose: The "Measure" phase quantifies and evaluates AI risks to inform decision-making.
    - Quantification Methods:
        - Risk Metrics: Develop metrics to measure risk severity, likelihood, and impact.
        - Scoring Models: Use scoring models to assess risks objectively.
        - Data Collection: Gather relevant data to support risk measurement.
    - Risk Prioritization: Prioritize risks based on their significance and potential consequences.
    - Risk Tolerance: Define acceptable risk thresholds.

- Manage: Implement risk mitigation strategies.

    - Purpose: The "Manage" phase involves implementing risk mitigation strategies.
    - Risk Treatment Options:
        - Avoidance: Eliminate or avoid high-risk AI practices.
        - Mitigation: Implement controls to reduce risk impact.
        - Transfer: Transfer risk through insurance or contracts.
        - Acceptance: Accept residual risk when other options are not feasible.

- Monitoring and Adaptation: Continuously monitor AI systems, adapt risk management strategies, and learn from incidents.
- Feedback Loop: Establish a feedback loop to improve risk management practices over time.

4. Roles and Responsibilities

- Clearly define roles for AI risk management.

    - Purpose: Clearly defining roles ensures that everyone understands their responsibilities in managing AI risks.
    - Key Aspects:
        - Risk Managers: These individuals oversee the entire risk management process. They coordinate risk assessments, monitor risks, and ensure compliance with risk management policies.
        - Technical Experts: Technical experts (such as data scientists, AI engineers, and cybersecurity specialists) play a crucial role in assessing AI risks. They understand the technical aspects of AI systems and identify potential vulnerabilities.
        - Business Leaders: Business leaders (including executives, managers, and project owners) need to be aware of AI risks and make informed decisions based on risk assessments.
        - Legal and Compliance Teams: These teams ensure that AI systems adhere to legal requirements, industry standards, and organizational policies.
        - Communication Channels: Establish clear communication channels between these roles to share risk-related information effectively.

- Involve stakeholders across the organization.

    - Collaboration: Involve stakeholders from various departments (such as IT, legal, compliance, and business units).
    - Holistic View: Different stakeholders bring diverse perspectives, ensuring a holistic view of AI risks.
    - Risk Identification: Stakeholders contribute to identifying risks specific to their areas of expertise.

- Foster collaboration between technical and non-technical teams.

    - Bridge the Gap: Collaboration between technical (data scientists, engineers) and non-technical (business analysts, legal) teams is essential.

- Technical Insights: Non-technical teams benefit from technical insights to understand AI risks.
- Business Context: Technical teams need business context to assess risks effectively.
- Decision-Making: Joint collaboration ensures that risk management decisions align with organizational goals.

5. AI-specific Risk Factors

- Address unique AI challenges

    - Bias: AI systems can inherit biases from training data, leading to unfair or discriminatory outcomes. Addressing bias involves ensuring fairness across different demographic groups.
    - Explainability: AI models often operate as "black boxes," making it challenging to understand their decision-making process. Ensuring explainability allows users to comprehend why an AI system made a particular decision.
    - Robustness: AI systems should perform consistently across various scenarios and inputs. Robustness involves testing AI models against adversarial attacks, noisy data, and edge cases.

- Consider ethical, legal, and societal implications.

    - Ethical Considerations:
        - AI systems must align with ethical norms and principles.
        - Ethical guidelines address issues like privacy, consent, and transparency.

    - Legal Compliance:
        - AI systems must adhere to legal requirements (e.g., data protection laws, intellectual property rights).
        - Compliance ensures that AI deployment does not violate legal boundaries.

    - Societal Impact:
        - AI can significantly impact society, affecting employment, privacy, and social dynamics.
        - Responsible AI considers societal well-being and aims to minimize negative consequences.

Applying the NIST AI RMF in the Real World

Worldwide

- **IBM’s Approach to Implementing the NIST AI RMF**

  IBM, a renowned company in the technology industry, has demonstrated a strong commitment to AI ethics and responsible technology development. In line with this commitment, IBM has supported the development of the National Institute of Standards and Technology (NIST) AI Risk Management Framework (AI RMF) and has undertaken a thorough analysis to ensure alighnment with its internal standards and practices.
  IBM's three-phase approach to analyzing the NIST AI RMF involved studying and preparing, mapping to internal methodology, and conducting a systematic analysis. This process confirmed that IBM's internal standards, policies, and practice guidance align well with the framework. Key findings highlighted the importance of IBM's AI governence structure and Ethics Board in facilitating this alignment.
  The NIST AI RMF and associated playbook were found to be useful tools for organizations, guiding for identifying additional measures to address AI risks comprehensively. Recommendations were made for policymakers to incorporate the framework into regulations and for companies to implement it by identifing relevant documentation, mapping internal methodologies, and systematically evaluating alignment.
  IBM emphasizes the importance of continuously adapting best practices to match the evolving landscape of AI technology, encouraging other organizations to undertake similar efforts to manage risks and enhance the benefits of AI.

- **Comparing OpenAI's GPT-4 Safety Measures with NIST AI Risk Framework**

  OpenAI recently launched GPT-4, their most advanced AI model yet, which has significant potential but also raises concerns about safety and reliability. There's a growing debate about whether current methods for assessing AI risks are adequate. Some experts are calling for a pause in developing more powerful AI systems to establish stronger safety standards. The National Institute of Standards and Technology (NIST) has released guidelines for managing AI risks, offering a potential framework for regulation. This article examines OpenAI's safety testing of GPT-4, its limitations, and how it aligns with NIST's recomendations. It concludes with recomendations for improving AI risk management by Congress, NIST, industry labs, and funders.

What Did OpenAI Do Before Deploying GPT-4

1. They engaged over 50 experts, known as "red teamers," from diverse domains to thoroughly test the model for potential failures or harms, such as spreading misinformation, bias in outputs, aiding malicious activities, or exhibiting power-seeking behavior.

1. Identified issues found by red teamers were addressed through Reinforcement Learning on Human Feedback (RLHF), where human evaluators provided feedback on model outputs. This feedback helped refine the model by reinforcing desirable responses and reducing undesirable ones.

How Did OpenAI's Efforts Map Onto NIST's Risk Management Framework

OpenAI's efforts to ensure the safety of GPT-4 can be understood in the context of NIST's AI Risk Management Framework (RMF), which has four main functions: map, measure, manage, and govern.

Firstly, OpenAI mapped risks by identifying areas for investigation based on past harms caused by similar language models. Then, they measured these risks qualitatively through red-teaming efforts and quantitatively using internal evaluations.

To manage the risks, OpenAI relied on techniques like Reinforcement Learning on Human Feedback (RLHF) and shaping the dataset to improve GPT-4's behavior. However, NIST's framework lacks concrete standards for assessing the adequacy of these practices.

NIST recommends improving risk assessment and mitigation by providing more detailed guidance, concrete metrics, and benchmarks. Additionally, industry labs should share insights with standard-setters like NIST and incorporate more public feedback into their risk management processes.

Sri Lanka

After conducting a thorough search, we were unable to find any applications related to the NIST AI Risk Management Framework in Sri Lanka at the moment.

Reference List

1. *NIST*. (2024, April 9). NIST. https://www.nist.gov/
2. AI RMF 1.0 Handbook - https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf
3. AI RMF Playbook - https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook
4. Federation of American Scientists. (2023, May 11). *How do OpenAI's efforts to make GPT-4 "Safer" stack up against the NIST AI Risk Management Framework? - Federation of American Scientists*. https://fas.org/publication/how-do-openais-efforts-to-make-gpt-4-safer-stack-up-against-the-nist-ai-risk-management-framework/
5. *IBM's Approach to Implementing the NIST AI RMF - IBM Policy*. (2023, September 26). IBM Policy.https://www.ibm.com/policy/ibms-approach-to-implementing-the-nist-ai-rmf/
6. Hyperproof. (2024, March 21). *NIST AI Risk Management Framework Ultimate Guide*. https://hyperproof.io/navigating-the-nist-ai-risk-management-framework/