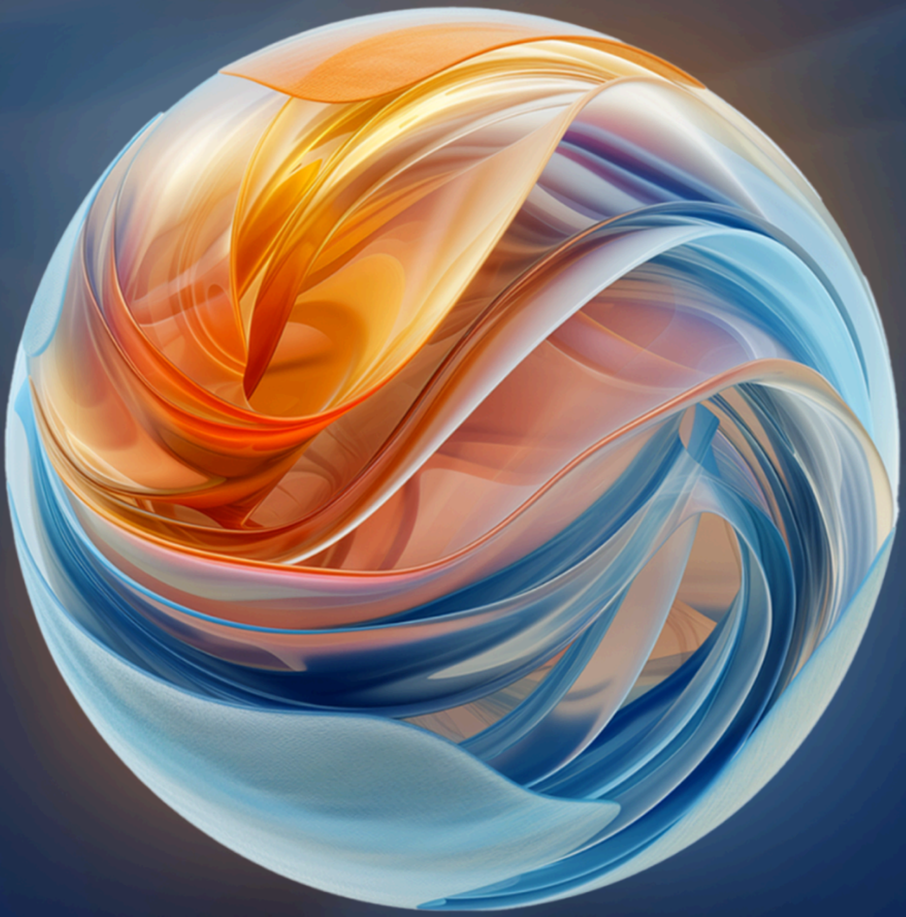
AI模型风险管理框架



AI Technology and Risk Working Group

CSA GCR cloud security alliance®
GREATER CHINA REGION

CSA cloud security alliance®

Table of Contents

The permanent and official location for the AI Technology and Risk Working Group is https://cloudsecurityalliance.org/research/working-groups/ai-technology-and-risk.

Acknowledgments

Lead Authors

Maria Schwenger
Vani Mittal

Contributors

Eric Tierling
Hadir Labib
Michael Roza
Renata Budko

Reviewers

Candy Alexander
Daniel C
Elier Cruz
Harie Srinivasa Bangalore Ram Thilak
Karan Goenka
Kenneth Thomas Moras
Namal Kulathunga
Nicolas Ray
Otto Sulin
Rohit Valia
Sanitra Angram
Tom Bowyer
Vaibhav Malik
Yuvaraj Madheswaran

Co-Chairs

Chris Kirschke
Mark Yanalitis

CSA Global Staff

Josh Buker
Marina Bregkou
Stephen Smith

Executive Summary

The widespread adoption of sophisticated machine learning (ML) models presents exciting opportunities in fields like predictive maintenance, fraud detection, personalized medicine, autonomous vehicles, and smart supply chain management [1]. While these models hold the potential to unlock significant innovation and drive efficiency, their increasing use also introduces inherent risks, specifically those stemming from the models themselves. Unmitigated model risks can lead to substantial financial losses, regulatory issues, and reputational harm. To address these concerns, we need a proactive approach to risk management. Model Risk Management (MRM) is a key factor for fostering a culture of responsibility and trust in developing, deploying, and using artificial intelligence (AI) and ML models, enabling organizations to harness their full potential while minimizing risks.

This paper explores the importance of MRM in ensuring the responsible development, deployment, and use of AI models. It caters to a broad audience with a shared interest in this topic, including practitioners directly involved in AI development and business and compliance leaders focusing on AI governance.

The paper highlights the inherent risks associated with AI models, such as data biases, factual inaccuracies or irrelevancies (known colloquially as “hallucinations” or “fabrications”[2]), and potential misuse. It emphasizes the need for a proactive approach to ensure a comprehensive MRM framework. This framework is built on four interconnected pillars: Model Cards, Data Sheets, Risk Cards, and Scenario Planning. These pillars work together to identify and mitigate risks and improve model development and risk management through a continuous feedback loop. Specifically, Model Cards and Data Sheets inform risk assessments, and Risk Cards guide Scenario Planning. Scenario Planning refines risk management and model development.

By implementing this framework, organizations can ensure the safe and beneficial use of ML models with key benefits such as:

- Enhanced transparency and explainability
- Proactive risk mitigation and “security by design”
- Informed decision-making
- Trust-building with stakeholders and regulators

This paper emphasizes the importance of MRM for harnessing the full potential of AI and ML while minimizing risks.

[1] McKinsey & Company “The state of AI in 2023: Generative AI’s breakout year”
[2] NIST AI 600-1 “Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile”

Intended Audience

AI MRM is designed for a broad audience with a common interest in the responsible development and deployment of ML models. It bridges the gap between technical and non-technical stakeholders, catering to those directly involved in the technical aspects of AI development and those concerned with its governance and oversight.

The target audience can be segmented into the following two primary groups.

1. **Practitioners in AI Model Development and Implementation**

   - **ML Engineers and Data Scientists:** This group benefits from the detailed explanations of Model Cards and Data Sheets and how they contribute to model understanding and development. Understanding these components empowers them to build more transparent and accountable models.

   - **AI Developers and Project Managers:** This group will find tools to help them anticipate potential issues throughout the AI model lifecycle, ensuring responsible deployment from conception to implementation.

2. **Stakeholders in AI Governance and Oversight**

   - **Risk Management Professionals, Compliance Officers, and Auditors:** This group will find sections on the importance of MRM and its alignment with common industry frameworks most relevant to establishing, enforcing, and assessing effective governance practices.

   - **Business Leaders and Executives:** The introduction and conclusion sections will be particularly valuable for them, as they highlight the importance of MRM in fostering responsible AI adoption within the organization.

   - **Communications and Public Relations Professionals:** This group will benefit from sections on communicating AI model risks and benefits, stakeholder engagement, and reputation management, as well as learning how to craft resonant messages for diverse audiences.

Scope

This paper explores MRM and its importance for responsible AI development. It examines closely the four pillars of an effective MRM framework and how they work together to create a holistic approach to MRM. We discuss how these techniques foster transparency, accountability, and responsible AI development.

The paper emphasizes the role of MRM in shaping the future of ethical and responsible AI. Note that this paper focuses on the conceptual and methodological aspects of MRM, and does not address the people-centric aspects, such as roles, ownership, RACI, and cross-functional involvement, which are covered in the CSA publication "AI Organizational Responsibilities - Core Security Responsibilities".

Introduction

The Need and Importance of MRM

Today, we witness the adoption of complex AI/ML models at an unprecedented rate across diverse industries. On the one hand, the increasing reliance on ML models holds the promise of unlocking vast potential for innovation and efficiency gains. On the other hand, it introduces inherent risks, particularly those associated with the models themselves – model risks. If left unchecked, they can lead to significant financial losses, regulatory sanctions, and reputational damage. Biases in training data, factual inaccuracies in model outputs (called "hallucinations" or "fabrications"[3]), and the potential for misuse, alongside privacy risks and intellectual property (IP) concerns, necessitate a proactive approach to risk management. AI MRM emerges as a critical discipline to ensure the responsible and trustworthy development, deployment, and utilization of these models.

MRM is a term commonly used in industries like finance, where it traditionally refers to managing risks associated with quantitative models. In this paper, however, this established concept outlines a framework for managing the risks associated with AI models.

AI MRM helps safeguard against the complexities, uncertainties, and vulnerabilities associated with AI models, bolstering confidence among users, stakeholders, and regulators in the dependability and fairness of AI-driven decisions. As AI continues to evolve and permeate more sectors, MRM will play an increasingly vital role in shaping the future of responsible AI deployment, benefiting businesses and industries.

At its core, model risks arise from the inherent limitations of the models themselves. Several of the most frequently seen sources of AI model risks are:

- **Data Quality Issues:** The foundation of any model is its data. Inaccurate, incomplete, or biased data can lead to a flawed model, resulting in unreliable outputs and erroneous conclusions. For example, if a model is trained to predict loan defaults using historical data that underrepresents high-risk borrowers, it might underestimate the risk of future defaults, potentially leading to financial losses.

[3] NIST AI 600-1 "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile"

- **Model Selection, Tuning, and Design Flaws:** Choosing the wrong model architecture or employing inappropriate algorithms for the given task can significantly impact the model's effectiveness and reliability. For instance, using a linear regression model to predict a highly non-linear phenomenon like stock market volatility would likely yield misleading results. It is also important to ensure the model's integrity, particularly when using open-source models, as end users should be able to verify the model's signature to ensure they are using the correct model and that the Model Card accurately represents the model's capabilities and limitations.

- **Inherent Risks to the Best-in-Class Models**: Even top-performing models released by big-name vendors may carry intrinsic risks based on the shortcomings of the model themselves, such as hallucination, harmful language, bias, and data leakage. These risks can have far-reaching impacts, affecting not only individual organizations but also society as a whole[4].

- **Implementation and Operational Errors:** A well-designed model can be compromised during implementation. Incorrect coding, inadequate controls, or improper integration with existing systems can introduce model deployment errors. For instance, a credit scoring model may be correctly developed, but its implementation into the loan processing system may be flawed, leading to inaccurate assessments and unfair loan denials. Security is another key operational set of risks. Those risks can be both well-established, such as application-level and access-level vulnerabilities, and new in the GenAI era, such as prompt injections[5]. AI models also increase the risks of threat actors aiming to change the decision-making that model users aim to use the model for.

- **Evolving External Factors:** Models are often trained on historical data, assuming a certain level of stability in the underlying environment. However, the real world is constantly evolving. Economic downturns, new regulations, or unforeseen events can render historical data irrelevant, leading to unreliable predictions by the model. For example, a model trained to predict customer churn based on past purchasing habits might struggle if consumer preferences shift due to a global pandemic. Similarly, a model trained to predict loan defaults based on historical data may also struggle if consumer behavior shifts due to unforeseen events such as a global pandemic, changes in economic policies, or unexpected changes in loan activity (new, refinance, and renegotiation of terms). Both examples illustrate how models can be vulnerable to unexpected changes in the underlying environment, highlighting the need to monitor and update models to ensure their effectiveness.

An MRM framework is a structured approach to identifying, assessing, mitigating, and monitoring the risks associated with ML models, especially in decision-making processes. Establishing this is a proactive practice that safeguards the benefits of ML models while minimizing potential downsides. It acts as a roadmap for organizations to ensure the responsible and trustworthy development, deployment, and use of these models. It's important to note that the specific risks and their severity (risk level) will vary depending on the nature of the organization, industry, business unit, and the model's intended use.

[4] CNBC The biggest risk corporations see in gen AI usage isn't hallucinations
[5] World Wide Technology Secure Your Future: A CISO's Guide to AI

A well-designed MRM framework enables customization by establishing a structured process to identify and assess model-specific risks. This ongoing process is built on several essential components, which include the following:

1. Governance

The governance of AI and ML models within an organization is critical for ensuring that these models are effectively managed and aligned with the strategic goals and regulatory requirements. This involves setting clear objectives, maintaining a detailed inventory, defining ownership roles, and establishing approval processes. Key components of governance include:

- **Business Approach:** Defines the organization's overall AI strategy and business goals to identify areas where AI can be leveraged to improve productivity, efficiency, decision-making, or deliver new user experiences.
- **Model Inventory:** Establishes a comprehensive list of all models used within the organization, categorizing them by purpose, complexity, risk level, and alignment with the established business approach. A well-structured model inventory enables targeted risk assessment and monitoring of high-risk or critical models through categorization based on risk level and potential impact.
- **Model Lifecycle Management:** Clearly define roles and responsibilities for each model's life cycle, from design and testing to development and deployment, ongoing monitoring and maintenance, and deprecation. Clear ownership enables efficient knowledge transfer and documentation, reducing the risk of knowledge gaps or silos that could hinder the model's long-term maintenance and evolution.
- **Model Approvals:** Establishes a formal process and criteria for approving models before deployment, ensuring they meet business needs, align with the business architecture, and comply with regulatory requirements. The approval process also evaluates models for potential biases, ethical concerns, and adherence to responsible AI principles, promoting fairness, transparency, and trustworthiness.

2. Model Development Standards

Establishing robust model development standards is essential for ensuring that AI models are built on high-quality data, adhere to best practices and comply with relevant regulations. This includes managing data quality, following standardized design and development practices, and implementing thorough validation and testing processes. Key components of model development standards include:

- **Data Quality Management:** Defines practices to ensure high-quality data for model training by requiring accuracy, completeness, minimal bias, and minimization (ensuring data is fit for the purpose and limited to only necessary information) through data diversification and adherence to intellectual property and privacy protection measures.

- **Model Design and Development:** Outlines standards for model architecture, development methodologies, and documentation practices. Align model development standards with existing governance and compliance frameworks, including regulatory guidelines. For a list of the most prominent guidance, see "Appendix 1: AI Frameworks, Regulations, and Guidance."

- **Model Validation and Testing:** Establishes processes for rigorously testing models to assess their performance, accuracy, safety and robustness.

- **Governance and Compliance Frameworks:** Align model development standards with existing governance and compliance frameworks, including recommendations by regulatory guidelines (e.g., GDPR, CCPA), industry standards (e.g., ISO 27001, ISO 42001), and organizational policies. For guidance on ensuring adherence to legal, ethical, and risk management requirements, refer to the CSA publication "Principles to Practice: Responsible AI in a Dynamic Regulatory Environment".

3. Model Deployment and Use

- **Model Monitoring:** Implements procedures for continuously monitoring model performance in production, detecting any degradation in accuracy or unexpected behavior.

- **Model Change Management:** Defines a transparent process for changing deployed models, ensuring proper testing and validation before implementation, and providing rollback and deprecation mechanisms for models no longer in use.

- **Model Communication and Training:** Establishes protocols for communicating model limitations and capabilities to stakeholders and providing training for proper model usage.

4. Model Risk Assessment

Model risk assessment is essential for identifying and mitigating potential risks in AI and ML models, both developed internally and acquired externally. This process addresses risks across financial, supply chain, legal, regulatory, and customer domains. Key components include:

- **Risk Scope:** Risk assessment process applies not only to models developed and used within an organization, but also to models acquired from third parties or outside organizations. It defines the types of risks the organization would like to address at all levels, such as financial, supply chain, legal and regulatory, customer retention, and so on.

- **Risk Identification:** This is the initial step in effectively managing risks associated with ML models. It involves employing techniques to uncover potential issues throughout the model lifecycle systematically. Some key factors considered during risk identification include data quality, model complexity, intended use, training data acquisition and use of personal data, and model protection mechanisms.

- **Risk Assessment:** Evaluate the severity and likelihood of identified risks, allowing for prioritization of mitigation efforts. Risk assessment may use qualitative or quantitative methods, such as FAIR-AI[6].

- **Risk Mitigation:** Develop strategies to address the identified risks, including data cleansing, model improvements, implementing security and privacy controls, and protecting intellectual property. Prioritize the efforts based on the balance of risk reduction of those efforts against their costs and practicability within the organization's environment.

5. Documentation and Reporting

Thorough documentation and regular reporting are vital for maintaining transparency and accountability in model risk management. These practices ensure that all aspects of the model lifecycle are well-documented and communicated to relevant stakeholders. Key components include:

- **Model Documentation:** Maintains comprehensive documentation throughout the model lifecycle, capturing development steps, assumptions, limitations, and performance metrics.

- **Model Risk Reporting:** Regularly reports to relevant stakeholders on identified model risks, mitigation strategies, and overall model performance.

A robust MRM framework ensures trustworthy development, deployment, and ongoing use of ML models. By proactively identifying, assessing, and mitigating these risks, organizations can harness the power of models while safeguarding themselves and their customers and users from potential pitfalls. This ensures the reliability and accuracy of model-driven decisions and fosters trust and transparency.

[6] A FAIR Artificial Intelligence (AI) Cyber Risk Playbook

The Four Pillars: Model Cards, Data Sheets, Risk Cards, Scenario Planning

This framework can be built by combining four key components:

- **Model Cards:** Offer a clear and concise window into an ML model. They detail the model's purpose, training data, capabilities, Adversarial AI resistance, limitations, and performance, promoting transparency and informed use.
- **Data Sheets:** Function as detailed descriptions of a dataset used to train an ML model. They document the creation process, composition (data types, formats), intended uses, potential biases, limitations, and any ethical considerations associated with the data.
- **Risk Cards:** Summarizes the key risks associated with an AI model. It systematically identifies, categorizes, and analyzes potential issues, highlights observed risks during development or deployment, explains current and planned remediations, and outlines expected user behavior to ensure responsible use of the model.
- **Scenario Planning:** Explores hypothetical situations where a model could be misused or malfunctioning, helping identify unforeseen risks and develop mitigation strategies.

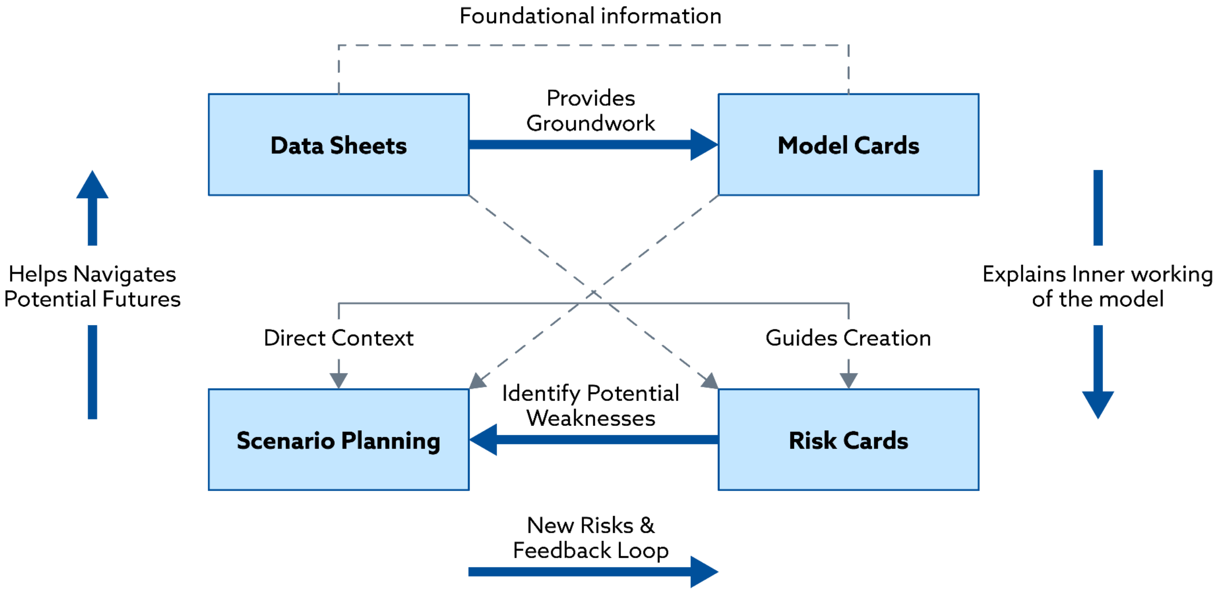

Foundational information
Data Sheets
Provides Groundwork
Model Cards
Helps Navigates Potential Futures
Explains Inner working of the model
Direct Context
Guides Creation
Scenario Planning
Identify Potential Weaknesses
Risk Cards
New Risks & Feedback Loop

Figure 1. Framework Pillars for Responsible and Well-Informed Use of AI/ML

These techniques work together to create a holistic approach. In a nutshell, information from Model Cards informs risk assessments; building on the foundation of Model Cards and Data Sheets provide additional context for understanding model strengths and limitations. Risk Cards guide Scenario Planning exercises, and Scenario Planning outcomes feed back into risk management, creating a continuous feedback loop.

Note: *The difference between the Training Data category in Model Cards and the Technical Specifications section in Data Sheets is that the Training Data category in Model Cards refers to the specific dataset used to train a machine learning model, including its sources, size, quality, and preprocessing steps. On the other hand, the Technical Specifications section in Data Sheets provides a detailed description of a dataset's technical construction and operational characteristics, including data schema, processing steps, and technical dependencies, which are not limited to model architecture. Understanding this distinction is essential to effectively utilize both Model Cards and Data Sheets for managing and maintaining machine learning models and datasets.*

By combining these techniques, organizations can create a comprehensive Risk Management Framework (RMF) that fosters:

- **Transparency and Explainability:** Model Cards, Data Sheets, and clear communication empower stakeholders to understand model capabilities and limitations. Techniques like Local Interpretable Model-agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP), Integrated Gradients, Concept Activation Vectors (CAVs), and model distillation can provide local explanations, identify high-level semantic concepts, and create interpretable surrogate models, respectively, to enhance transparency and explainability of complex models.

- **Proactive Risk Management:** A multi-faceted approach is key to effective risk management. This includes utilizing Model Cards to document potential biases and limitations, leveraging Data Sheets to understand the training data, conducting thorough risk assessments (based on the Risk Cards) to identify general risks, and engaging in Scenario Planning to explore potential future challenges. Additionally, adversarial testing, stress testing, edge-case analysis, and regularization techniques like dropout, L1/L2 regularization, or adversarial training can help identify vulnerabilities, blind spots, and improve model robustness, enabling proactive risk management.

- **Consistent Risk Management:** Consistency in the risk management process is focused on ensuring that risk assessments are reproducible and allow comparison and tracking of AI models' performance and safety over time. Consistent risk assessments help accurately monitor risks' evolution and mitigation strategies' effectiveness, fostering continuous improvement in AI systems.

- **Informed Decision-Making:** A comprehensive understanding of model risks empowers stakeholders to make informed decisions about model deployment and use.

- **Building Trust, Credibility, and Ethical Use:** Transparency and responsible risk management practices build trust and foster ethical use of ML models. Implementing privacy-preserving techniques, obtaining certifications for ethical AI practices, establishing governance frameworks and ethical AI committees, and conducting third-party audits can build trust, credibility, and foster ethical use of ML models.

- **Continuous Monitoring and Improvement:** Ongoing monitoring and the ability to adapt based on new information ensures the model's sustained effectiveness and safety. Some techniques include adopting Machine Learning, Security, and Operations (MLSecOps) practices. This involves setting up monitoring pipelines to track model performance, data drift, feedback loops, and unintended consequences. Additionally, implementing online or continuous learning techniques is important. Establishing processes for incorporating user feedback, incident reports, and lessons learned can ensure sustained effectiveness, safety, and continuous improvement of AI systems.

Benefits of a Comprehensive Framework

A comprehensive risk management framework (RMF) for ML models offers several benefits, defined below.

Enhanced Transparency, Explainability, and Accountability

Model Cards, Data Sheets, and Risk Cards are pivotal for transparency, explainability, and accountability within MRM. Data Sheets document the data's origin, acquisition composition, and pre-processing methods, providing crucial context for understanding a model's inputs, limitations, and role. This documentation can help you understand the inner workings of a model to a degree, allowing for some assessment of its strengths, weaknesses, and potential bias. What is available for proprietary models is typically much more restricted when compared to open-source models.

Proactive Risk Assessment and Scenario Analysis

Data Sheets complement Scenario Planning by detailing the data-specific characteristics that could influence model performance under different scenarios. This information is vital for conducting thorough risk assessments and ensures that scenario analyses consider data quality and other factors relevant to the company.

Development of Risk Mitigation Strategies

Incorporating insights from Data Sheets into the risk mitigation process allows for more targeted strategies. Understanding data limitations and biases helps in designing effective mitigations, such as data cleansing, augmentation, or re-balancing techniques, which are critical for addressing potential risks identified by Risk Cards.

Informed Decision-Making and Model Governance

Data Sheets, which detail the training data and model characteristics, are critical in informing governance practices. This detailed understanding ensures well-founded, documented, and transparent decisions around model deployment. While training data can be swapped out, its quality directly impacts the model's behavior. Data Sheets help identify potential limitations and biases within the data, as well as the characteristics of the model that might influence its outputs. This comprehensive information leads to informed decisions about model deployment. Data Sheets provide essential information that influences decision-making processes by highlighting data-related constraints and opportunities. In MRM, this detailed understanding of data characteristics informs governance practices, ensuring decisions around model deployment are well-founded, documented, and justifiable.

Robust Model Validation

Robust model validation is integral to the MRM framework, ensuring that models perform as expected and adapt to real-world conditions. This involves rigorous testing with diverse datasets that reflect real-world scenarios. Information from Data Sheets, such as the data distribution and potential biases, can inform the selection of these datasets for a more comprehensive validation process. Techniques like diversity testing, stress testing, and generalizability metrics are crucial for this validation process. By incorporating these validations, the framework ensures that models maintain effectiveness and avoid unexpected performance issues or biased outcomes in real-world applications.

Building Trust and Enhancing Model Adoption

Data Sheets form the groundwork for trust by ensuring data clarity. However, building confidence requires a multi-layered approach. Model Cards provide deeper insights into the model's inner workings, and Risk Cards proactively address potential biases or limitations. This promotes transparency and responsible AI development, ultimately leading to increased user and regulatory trust in adopting the model. These documents provide transparent and honest communication about model capabilities and performance expectations. This clarity is crucial for gaining the confidence of users and regulators, particularly in sectors where data provenance and integrity are critical.

Continuous Monitoring and Improvement

Continuous monitoring is integral to the MRM framework, ensuring that models operate as expected and adapt to changes over time. This involves regular updates to Model Cards, Risk Cards, and Data Sheets to reflect changes in the model's performance or operational environment. For example, metrics such as accuracy, precision, and F1 score may be tracked to measure performance, while Mean Absolute Error (MAE) and Mean Squared Error (MSE) may be used to evaluate model drift. Continuous monitoring helps identify when a model might be drifting from its intended performance or when changes in the external environment necessitate model adjustments or deployment strategy. This ongoing vigilance helps ensure sustained compliance, efficacy, and safety of ML models in dynamic operational contexts.

Positive Societal and Ethical Impact

Data Sheets are foundational for addressing societal and ethical bias in ML models. Documenting the training data's origin, composition, and preprocessing methods provides crucial transparency to identify potential biases, which is crucial for developing fair and equitable ML models. By ensuring that data handling practices are aligned with ethical standards, organizations can better manage the broader impacts of their technologies.

Strong Governance and Oversight

Strong governance and oversight, built on a foundation of controls that ensure alignment with the organization's objectives, guarantee transparent, explainable, and accountable AI model development, use, and maintenance, guided by ethically aware and competent individuals. They establish robust enforcement mechanisms that ensure ethical guidelines and responsible data practices are followed. Effective governance involves clear roles and responsibilities, defined decision-making processes, and escalation procedures for resolving conflicts or disputes. Regular audits provide a layer of accountability, verifying stakeholder commitment to these principles. Rigorous change management procedures, control updates, retraining, and deployment decisions promote oversight and proactively mitigate potential risks. Clear communication and collaboration among stakeholders, including users, data scientists, engineers, and business leaders, are crucial for successful governance and oversight.

Core Components

1. Model Cards: Understanding the Model

Model Cards provide a transparent overview of a model. They detail the model's purpose, training data, capabilities, limitations, and performance metrics. This information helps developers, deployers, risk management professionals, compliance officers, and end-users understand the model's strengths and weaknesses, forming the foundation for risk assessment.

Key elements of a Model Card usually include:

- **Model's Details and Intended Purpose:** This clarifies the model's function and goal.
- **Training Data Details:** This describes the composition of the data used to train the model, including its source, size, how it was acquired (consent, donation, etc.), ethical considerations, and potential biases. A link to a Data Sheet (if available) can be provided for further details.
- **Intended Use Cases and Limitations:** This explains what the model can be used for and where it might not perform well.
- **Performance Metrics (Evaluation Metrics):** This outlines how well the model performs on relevant tasks, using clear metrics like accuracy and generalizability.
- **Evaluation Methodologies Employed:** This describes the methods used to assess the model's performance.
- **Model Explainability and Bias:** This section describes techniques for understanding the model's decision-making process and identifying potential biases. It also details methods for mitigating bias and ensuring fair outcomes across different groups.
- **Known Limitations:** This acknowledges potential shortcomings of the model, such as susceptibility to specific prompts or factual errors.
- **Sustainability and Environmental Aspects (Optional):** If available, this estimates the environmental impact of training the model (e.g., carbon emissions).
- **Adversarial Resistance (Performance Metrics under Adversarial Attack–Optional):** Although specific details of adversarial training are not usually documented in the Model Cards, based on our experience, we recommend including adversarial resistance metrics in the evaluation section of the Model and Risk Cards. Data scientists can demonstrate a model's resilience by reporting accuracy metrics under simulated adversarial attacks, providing a more comprehensive understanding of the model's performance and potential vulnerabilities.

Benefits of Model Cards

Model Cards offer a wealth of advantages that contribute to responsible AI development and deployment and serve as a foundation of risk management, including:

- **Insights and Transparency:** Model Cards guide stakeholders, helping them understand the model's design, development process, and deployment. They illuminate the training data and the model's performance metrics, allowing users to grasp its capabilities and limitations.
- **Identifying Potential Risks:** By outlining the composition of the training data, Model Cards can reveal potential issues like bias when the outputs may be influenced in unfair or discriminatory ways, copyright violations, limited generalizability when the model might not perform well in contexts different from its training data, factual errors stemming from inaccuracies in the training data, and others.
- **Reproducibility/Accountability:** Model Cards document the development process, enabling others to recreate the model and independently assess its risks.

Foundation for Risk Management

Model Cards serve as the cornerstone for effective risk management of ML models, providing key information about a model, including:

- **Training Data Characteristics:** Revealing potential privacy breaches, copyright infringement, and biases.
- **Behavior and Performance Limitations:** Anticipating situations where models might generate unreliable or misleading outputs.

Benefits for Risk Mitigation

- **Tailored Mitigation Strategies:** Knowing the types of risks allows seeking relevant mitigation strategies and then focusing on the ones with the highest risk reduction potential at acceptable implementation complexity, for example, developing specific safeguards against risks like generating harmful content
- **Communication and Transparency:** Facilitating stakeholder communication and responsible use
- **Guiding Prompt Design:** Designing prompts for safe and accurate responses

- **Compliance and Trust:** Assessing compliance with regulations, fostering trust, and ensuring informed decisions about model trustworthiness and safety
- **Training Data Curation:** Ensuring data quality and fairness
- **Implementing Guardrails:** Documenting techniques to prevent unintended outputs

In essence, Model Cards act as a comprehensive record, promoting responsible AI development and deployment and establishing the foundation for risk management and mitigation.

Creating and Updating Model Cards

Model Card Creation Essentials

Effective Model Card creation requires a collaborative and automated approach to ensure accuracy and efficiency. The most common best practices include the following:

- **Process and Ownership:** A clear process and ownership for creating and maintaining Model Cards must be established within the organization. Key leaders are responsible for enforcing this process and appointing a specific owner for each Model Card. The selected owners should have the skills to ask the right questions, gather necessary information, and lead collaborations across the organization. Ideally, they should have experience building Model Cards or be able to learn quickly, with sufficient technical knowledge.
    - Not every model may require a Model Card, so clear guidelines should be defined for when Model Cards are necessary, for example, for models used by over 100 people or in production or testing.
- **Collaboration:** Involve cross-functional teams in the creation process to ensure comprehensive coverage.
- **Template:** Use a standardized template to ensure consistency and ease of use.
- **Automation:** Leverage automation tools to generate Model Cards, reducing manual effort and increasing accuracy.
- **Version Control:** Utilize version control systems to track changes and maintain a clear record of updates.
- **Model Card Repositories:** Establish a centralized repository for Model Cards, ensuring easy access and management.

Keeping Model Cards Up to Date

Regular updates are critical to ensure Model Cards remain accurate and relevant. Implementing a streamlined update process reduces manual effort, increases efficiency, and should include:

- **Regular Reviews:** Conduct regular reviews of Model Cards to reflect changes in the model or data.
- **Automated Updates:** Utilize automation tools to update Model Cards, reducing manual effort and increasing accuracy.
- **Change Management:** Establish a process to document and approve updates properly.
- **Audit Trail:** Maintain an audit trail of all updates and changes to ensure transparency and accountability.

Some additional advanced techniques can be leveraged to create a streamlined and efficient process for creating and updating Model Cards. For example, ML algorithms can analyze model performance and update Model Cards dynamically, while natural language processing algorithms can generate Model Card content automatically. Visualization tools can provide a graphical representation of model performance and updates, making complex data easier to understand. Integrating Model Cards with other tools and systems, such as version control and collaboration platforms, can enhance collaboration and reduce manual effort. These approaches can improve the process with enhanced accuracy, efficiency, and collaboration.

Limitations of Model Cards

- **Completeness and Accuracy:** The details rely solely on how thoroughly and accurately the Model Card is filled out. This leaves a risk of misleading or incomplete information, especially when this process is primarily manual. For this reason, we advocate automating data collection as much as possible. However, ensuring completeness and accuracy also requires a cultural shift within the organization, sponsored and enforced by management, to prioritize Model Card updates and maintenance. Without leadership buy-in, even well-intentioned developers may deprioritize Model Card creation and updates, hindering the effectiveness of this risk management tool.
- **Static Representation**: Model Cards offer a valuable snapshot of a model at a specific time, but their static nature can pose challenges. As models are updated and improved, the information documented in the Model Cards may become outdated. This necessitates regular review and updates to the Model Card to ensure it accurately reflects the model's current state.
- **Subjectivity in Evaluation**: Models focusing on fairness or ethical consideration can be inherently subjective as no standardized benchmarks or evaluation criteria exist.

- **Limited Scope:** While Model Cards provide technical details like architecture, training data, and performance metrics, they often fall short of comprehensively addressing the model's impact. This limited scope can overlook potential biases, ethical considerations, and social implications that arise from the model's real-world application.
- **Varying Levels of Detail:** There's no standardized format for Model Cards. The level of detail and clarity can vary, making comparisons and risk assessment across different models difficult.

Model Cards are valuable tools for understanding ML models and their potential risks. They promote transparency and allow developers and users to understand the model's strengths and weaknesses.

2. Data Sheets: Examining the Training Data

Data Sheets of model blueprints provide an in-depth technical description of an ML model. They serve as a reference document for developers, risk managers, and auditors, detailing the model's construction parameters and operational characteristics. This information is crucial for understanding the model's potential strengths, weaknesses, and inherent risks.

The Need for Data Sheets

While Model Cards and Risk Cards offer valuable insights for risk management, an essential element still needs to be added: a transparent view of the model's internal logic. Data Sheets bridge this gap as a foundational document for effective model risk management. Here's how Data Sheets foster trust and enable more informed risk assessments:

- **Model Transparency:** Understanding how a model arrives at its decisions is crucial for risk management. While Model Cards provide a high-level overview and Risk Cards highlight potential issues, they don't delve into the model's inner workings. Data Sheets address this gap by looking deeper into the model's logic. This transparency fosters trust in the model and empowers risk managers to make more informed assessments of its limitations and potential biases.
- **Risk Assessment:** By understanding the model's construction and training data, risk managers can effectively evaluate potential sources of model risk, such as data quality issues, overfitting, or algorithm bias.
- **Model Governance:** Data specifications serve as a cornerstone for model governance practices, facilitating ongoing monitoring, maintenance, and retraining of the model as needed.
- **Reproducibility:** Detailed specifications ensure independent parties can recreate and validate the model, promoting trust and confidence in its outputs.

The Role of Data Sheets in MRM

Beyond simply documenting the model's logic, Data Sheets empower proactive risk management and ensure model fit. They provide the roadmap for ongoing improvement and compliance, fulfilling critical functions in the MRM lifecycle as follows:

- **Risk Identification and Mitigation:** Data specifications enable risk managers to proactively identify potential failure points within the model and develop mitigation strategies.
- **Model Validation and Refinement:** The documented training process and performance metrics allow for rigorous validation of the model's effectiveness and generalizability. Data specifications also provide a basis for ongoing calibration and refinement of the model to address identified biases or performance limitations.
- **Regulatory Compliance:** Comprehensive data specifications can play a vital role in demonstrating compliance with relevant regulations and ethical guidelines for AI/ML model development and deployment.

Key Elements of a Data Sheet

Data Sheets provide a concise and accessible overview of the model's inner workings, including:

- **Model Purpose and Scope:** Clear definition of what the model is designed to achieve and the limitations of its use.
- **Data Inputs and Assumptions:** A detailed listing of all input features the model uses, including data sources/types/formats and any pre-processing transformation steps applied, together with any underlying assumptions made.
- **Model Architecture:** A technical description of the model's architecture (e.g., decision tree, neural network), including hyperparameter settings (learning rate, number of layers) and the chosen algorithm.
- **Model Development Process:** Briefly outline the steps to build and train the model, including any relevant algorithms used.
    - **Training Data Characteristics:** A breakdown of the training data used to develop the model, encompassing data source(s), size, distribution characteristics, and any data quality checks performed.
    - **Training Process:** Documentation of the training process, including the chosen optimization algorithm, objective for success, and convergence criteria.

    - **Performance Metrics:** This is a comprehensive set of indicators used to evaluate the model's effectiveness at the training and validation datasets (e.g., accuracy, precision, recall, F1 score).

- **Model Outputs and Interpretation:** A clear definition of the model's output format, including data types and how the interpretations of the generated results should be understood.
- **Assumptions and Limitations:** Transparent disclosure of any assumptions made during model development and any limitations inherent to the chosen model architecture or training data.

Limitations of Data Sheets

While Data Sheets offer significant advantages, it's crucial to acknowledge their limitations to ensure they are used effectively. Data Sheets can present challenges in complexity and scope and keep pace with the evolving field of AI/ML. Some of these limitations include:

- **Complexity:** Depending on the specific components of the AI/ML framework, including the training dataset, selected algorithm, machine learning operations (MLOps) control regime, and performance metrics measured, the data specifications can become highly technical, requiring ML expertise to comprehend fully.
- **Limited Scope:** Data specifications primarily focus on the technical aspects of the model. They may not fully capture the broader business context or potential societal implications of the model's outputs.
- **Evolving Field:** As AI/ML rapidly evolves, data specification best practices may need to be continually adapted to incorporate new technologies and methodologies.
- **Common Limitations with Model Cards**, such as completeness and accuracy, becoming a company culture, and static/outdated representation, also apply to Data Sheets.

Data Sheets are an essential tool for managing model risk. By providing a technical roadmap for the model's construction and operation, they empower risk management professionals to effectively assess, mitigate, and govern the risks associated with ML models.

3. Risk Cards: Identifying Potential Issues

Risk Cards delve deeper into potential issues associated with AI models. They systematically identify, categorize, and analyze potential risks. Think of them like flashcards for potential model risks. Each card describes a specific risk, potential impact, and mitigation strategies. Similar to flashcards, they provide a quick and structured way to understand and address model vulnerabilities.

Risk Cards typically encompass a range of potential concerns, including:

- **Safety and Ethical Risks:** These encompass issues such as privacy, generation of harmful content, and the promotion of bias.
- **Security Risks:** Data breaches, manipulation attempts, and other security vulnerabilities fall under this category.
- **Societal Risks:** Job displacement or the misuse of AI for propaganda are examples of societal risks.
- **Environmental Risks**: AI models can use a lot of electric power, and thus increase the generation of harmful gasses. Even models that use clean energy take that energy away from other social uses, thus forcing them to generate harmful gasses.
- **Operational Risks:** Models can face challenges related to limited training data, compute intensity, integration with existing systems, and so on.
- **Regulatory and Legal Risks:** The organization may fall foul of laws, rules, regulations (LRR) due to its initial implementation or due to LRR changing over time. Or use of input data may be challenged by owners of intellectual property rights.
- **Financial Risks**: Costs of serving the model can increase unexpectedly, such as using agentic workflows.
- **Supply Chain Risks**: Relate to risks carried from outside the organization and ones with potential to carry from our model to partners.
- **Reputation Risks**: Inappropriate model usage can lead to negative press, and so on.

Note that risk categories may differ for your organization, or at least the depth of focus on each risk category may differ. For example, the NIST AI RMF[7] focuses on risks to a model being “valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed”.

Structure of Risk Cards

Each Risk Card follows a well-defined structure to ensure a focused and informative approach toward understanding the specific risk and developing targeted mitigation strategies. The following elements can be typically found in each Risk Card:

- **Risk Category:** Classify risks (e.g., bias, factual errors, misuse)

[7] NIST AI 100-1 “Artificial Intelligence Risk Management Framework (AI RMF 1.0)”

- **Risk Description:** A concise description of the potential problem, such as bias, factual errors, or generating harmful content
- **Impact:** The potential consequences of the risk, considering factors like reputational damage, user harm, or legal issues
- **Severity Level:** Assess the potential impact of the risk (high, medium, low)
- **Likelihood**: Evaluate the probability of the risk occurring
- **Mitigation Strategies:** Actionable steps to reduce the likelihood or severity of the risk could involve data filtering techniques, improved training data, user prompts guiding the model's development toward safer outputs, and operational and organizational strategies

The table below presents an example of a Risk Card.

Risk	Description	Impact	Mitigation Strategies
Bias & Fairness	Model outputs biased content based on training data	Promotes discrimination and a potential for reputational damage	▪ Use diverse training data ▪ Implement fairness checks in the model ▪ Provide transparency on limitations

This Risk Card highlighted the potential for unintended bias in the ML model used by a retail company for generating marketing and social media content. With a clear description and 'High' potential impact (high severity), the data team prioritized addressing the issue. The company conducted a dual review of the training data and the model architecture to investigate potential biases. The data team analyzed data demographics, identified skews in representation, and examined the sources of the training data for potential bias. They also discussed fairness metrics to quantify potential biases and used techniques like interpretability methods to understand how the model arrives at its outputs.

Based on this analysis, several mitigation strategies were implemented:

- **Data De-biasing:** Balancing the training data through oversampling/under-sampling and removing non-essential sensitive attributes were used to create a more balanced dataset. The company is also exploring using synthetic data to address bias further.
- **Fairness in Training:** Fairness constraints were incorporated into the training process to penalize biased outputs and reinforce appropriate outputs.
- **Post-processing Filters:** Deploying sentiment analysis and fact-checking tools to identify and flag potentially biased content after generation.

Beyond these mitigation strategies, the company also developed a well-thought-out contingency plan to reinforce the team's defense against bias. This contingency plan included:

- **Flag and Address Biased Outputs:** A process to clearly flag and address biased outputs involves human reviewers who can identify and correct biased content.
- **Incident Response Protocol:** When a Risk Card scenario is triggered, it is highly beneficial if the organizations already have a pre-established incident response protocol that can be leveraged by AI/ML Ops teams to ensure swift investigation and mitigation. Actions may include retraining the model with a more balanced dataset, as in the case of bias detection.
- **Communication Protocols:** Cross-company communication protocols, regarding potential bias, ensure transparency and foster trust with users and stakeholders, promoting responsible model usage throughout the organization.

By implementing these mitigation strategies, particularly the focus on data diversity and algorithmic fairness, the team took a proactive stance against bias in the model's outputs. This established a foundation for building trustworthy and ethical AI systems across the organization, enabling the company to promote inclusivity, transparency, and accountability in its AI applications.

Benefits of Risk Cards

Risk Cards offer a structured and dynamic approach to managing the ever-evolving landscape of model risk. They provide a systematic way to identify, categorize, and prioritize model risks and act as a powerful communication tool, facilitating discussions among developers, users, and stakeholders. This collaborative environment fosters a deeper understanding of potential issues, leading to the development of actionable insights such as mitigation strategies and contingency plans.

Beyond these core benefits, Risk Cards offer significant advantages specifically for MRM which include:

- **Proactive Approach:** Risk Cards help identify potential issues before they occur, allowing for preemptive solutions. This approach enables evaluating each strategy's potential risk reduction benefit versus its complexity and costs, ensuring proactive mitigation with the best return on investment.
- **Stress Testing:** Risk Cards facilitate the process of stress testing the model under various conditions by prompting discussions and brainstorming around potential risks. Risk Cards are a starting point for stress testing. Actual stress testing involves applying quantitative and qualitative techniques to analyze how the model would behave under those risks identified in the Risk Cards. The results of the stress tests are generally not recorded in the Risk Cards but may inform another iteration of the Risk Cards.

- **Improved Decision-Making:** Through comprehensive risk identification and analysis, Risk Cards empower organizations to make informed choices about deploying the model and selecting appropriate use cases. This ensures the model is utilized effectively while minimizing associated risks.

Limitations of Risk Cards

- **Limited Scope:** Risk Cards typically focus on a predefined set of potential issues. This can be beneficial for covering common risks, but it might not capture all the unique vulnerabilities specific to your AI model. This limitation also includes inadequate quantification, which hinders the assessment of risk impact and likelihood, making it challenging to prioritize mitigation efforts. Additionally, complex or nuanced risks might be oversimplified or condensed, which could lead to underestimating the severity or mitigation challenges.
- **Dynamic Nature of AI:** AI models constantly evolve, and new risks can emerge. Risk Cards need to be able to keep pace with the rapid development in the field.
- **Inadequate Quantification:** While Risk Cards provide a qualitative assessment of risks, they may fall short in quantifying the potential impact and likelihood of each risk. Without quantitative measures, organizations can struggle to prioritize and allocate resources effectively to mitigate the most significant risks associated with AI models.
- **Real-World Data Dependence:** The effectiveness of Risk Cards depends on the quality and comprehensiveness of the data used to identify and assess risks. Incomplete or inaccurate data might lead to misleading or irrelevant Risk Cards.
- **Human Judgment Required:** Risk Cards require human judgment to interpret the severity of a risk and choose appropriate mitigation strategies. This can be subjective and can depend on the expertise of the person reviewing the cards.

4. Scenario Planning: The “What If” Approach

Scenario Planning is a proactive approach exploring hypothetical situations where an AI model could be misused or malfunctioning. Essentially, it's about asking "what if". We imagine and explore how an AI model might behave in various positive and negative situations. This allows us to identify potential risks before they become reality.

Scenario Planning Considers

- Positive scenarios (e.g., increased productivity, improved education)
- Negative scenarios (e.g., weaponization of language, manipulation of information)

Aspects to Consider During Scenario Planning

- **Technical Capabilities:** Evaluate the model's strengths and weaknesses, focusing on areas susceptible to malfunctions (from regular to “black swan”[8]), manipulation or exploitation.
- **Data Biases:** Examine potential biases and data characteristics like less-trusted vendor data, missing or out-of-range data, and volatile-over-time data present in the training data that could influence the model’s outputs.
- **User Interaction:** Consider how users interact with the model and how their intent or understanding could lead to unintended consequences.
- **Societal Impact:** Explore potential broader societal impacts of model deployment, such as job displacement or ethical concerns surrounding automation or risks from usage of the model by people outside your organization.

How Scenario Planning Works

Scenario Planning involves a structured approach to identify and assess potential model risks through hypothetical situations. Here's a breakdown of the process:

1. Assemble the Team

Gather a diverse team with expertise in technology, risk management, ethics, legal, regulatory compliance, or specific data or application domains. The ideal team composition will depend on the project's specific requirements, and may include a combination of the following stakeholders:

- **Business Experts**
    - **Domain Experts:** Individuals who deeply understand the specific application domain (e.g., healthcare, finance) can provide valuable context for exploring scenarios relevant to real-world use cases.

[8] Wikipedia Black swan theory

    - **End-Users:** Including representatives of the intended user group provides insights into potential user interactions and how the model might be misused unintentionally.

- **Risk Experts**
    - **Security Practitioners:** Individuals with experience in threat modeling and quantifying the impact and likelihood of model vulnerabilities aid the risk discussion.
    - **Privacy and Legal Advisors:** Professionals with knowledge of the specific legal context of the organization and data being used, as well as privacy and information governance individuals can advise on privacy considerations for models processing personal data.
    - **Risk Management Specialists:** They bring experience in identifying and mitigating risks, ensuring a structured and comprehensive approach to Scenario Planning.
    - **Ethical Advisors:** Their expertise in ethical considerations helps to explore potential societal impacts and ensure responsible model development.

- **AI Experts**
    - **Model Developers:** Their expertise in model architecture and functionalities provides valuable insight into the system's capabilities and potential vulnerabilities.
    - **Data Scientists:** Their knowledge of the model’s training data and potential biases helps identify and estimate fairness and representation risks. Their knowledge of the model architecture clarifies the feasibility of management of specific risks.

By bringing together this diverse range of perspectives, the Scenario Planning team can better understand the AI model and identify a wider range of potential risks. This collaborative approach resembles product red-teaming, where diverse expertise and perspectives are leveraged to stress-test ideas and identify potential vulnerabilities. This approach also allows for blue-teaming capabilities, such as approaches for risk reduction. The effectiveness of this approach relies on assembling a team with the necessary bench strength to facilitate effective ideation and risk assessment.

2. Define Scope and Objectives

The next step involves clearly defining the scope and objectives of the Scenario Planning exercise. This includes specifying the AI system and the risks you want to explore. Establishing clear objectives, such as identifying potential biases, security vulnerabilities, or societal impacts, helps guide the team's focus and ensures a productive Scenario Planning session.

3. Prioritize Scenarios to Dive into

While a group that contributes diverse perspectives is great for proposing comprehensive potential scenarios, it can easily propose a list that is not feasible to plan through completely. That often necessitates careful prioritization. The team should pick their prioritization approach, such as some “t-shirt” sizes on definitions of “return” (e.g., potential risk impact vs. reduction) and “investment” (e.g., the effort Scenario Planning and implementation may take) for role comparisons. What matters is that the team prioritizes in a way that leadership feels comfortable with the risks of the scenarios that won’t be planned in as much detail.

4. Gather Information

The team should gather relevant information to understand the AI model and potential risks comprehensively. Model Cards, Data Sheets, and Risk Cards provide valuable insights into the ML models’ capabilities, limitations, and potential risks. These documents detail the training data, the model's architecture, and any known vulnerabilities. Additionally, researching relevant safety incidents or misuse cases involving the model helps the team anticipate potential real-world threats. The gathered information should have enough detail to plan a scenario, but no more than that.

5. Develop Scenarios

The core of Scenario Planning lies in creatively generating diverse hypothetical situations. Encourage the team to think outside the box and explore positive and negative scenarios. Techniques like "what-if" questions can spark creative thinking and create a broader range of scenarios. For example, the team might explore how a Large Language Model (LLM) used in customer service could be manipulated to generate biased responses or how a malfunctioning model in a financial setting could lead to inaccurate investment recommendations.

6. Evaluate Scenarios

Once scenarios have been developed, the team needs to analyze each one systematically. This involves considering the likelihood of the scenario occurring and the potential consequences, if it does materialize. The scenario's impact on various stakeholders, including users, society, and the organization, should be assessed. Consider how each scenario could affect the model's accuracy, reliability, fairness, and security. For instance, a scenario exploring the spread of misinformation by an LLM would need to consider the potential societal harm and reputational damage to the organization.

You can even use a language model to simulate these scenarios. Observe its outputs and identify potential risks, such as generating discriminatory text, spreading misinformation, or creating harmful content.

This step is among the most prone to scope creep (i.e., more work than initially budgeted for) and thus careful and disciplined project management is important here. Overly tight time control is also a risk. Ideally, this trade-off of evaluation depth versus coverage of key scenarios would be easier to manage with good upfront prioritization of scenarios.

7. Develop Mitigation Strategies

Based on the analysis of scenarios, formulate strategies to mitigate risks or adapt to future challenges. Develop contingency plans and response strategies to address potential scenarios that pose significant risks or threats to the organization. These strategies involve technical controls, such as implementing safeguards against manipulation, non-technical measures, such as user training on responsible model interaction, or enhancing transparency and accountability in AI governance processes. Additionally, adjustments to the model development process, like employing diverse training data sets, could be implemented to address potential biases.

8. Prioritize Mitigation Strategies to Implement

While a group that contributes diverse perspectives is great for proposing impactful mitigation strategies, the organization may not have the resources to implement all consistently. Thus, careful prioritization of strategies to implement will increase the probability that the key risks will actually be reduced. The team should pick their prioritization approach, as long as that gives confidence to the leadership team that all key risks are mitigated and that the deprioritized strategies indeed link to lower-probability and lower-impact risks.

9. Document and Communicate

The final step involves documenting the findings of the Scenario Planning exercise. This should include a comprehensive report outlining the scenarios explored, identified risks, the proposed mitigation strategies, and the recommended prioritization to implement. Sharing this report with relevant stakeholders, such as management, developers, and potential users, raises awareness of potential risks and guides decision-making throughout the model lifecycle. Effective communication fosters transparency and builds trust in the responsible development and deployment of AI models.

Benefits of Scenario Planning

- **Proactive Risk Identification and Mitigation:** Scenario Planning helps identify potential risks before they become reality, enabling timely mitigation efforts.
- **Improved Decision-Making:** By exploring various situations, stakeholders gain a more comprehensive understanding of model behavior, leading to better-informed decisions.
- **Enhanced Transparency and Trust:** Scenario Planning fosters open communication about potential risks, promoting transparency and building stakeholder trust.
- **Sustainable Model Development:** By testing models under various conditions, Scenario Planning helps identify weaknesses and informs improvements to make them more robust and reliable. This fosters continued responsible development and deployment of AI models.

Limitations of Scenario Planning

- **Limited Foresight:** The complexity of AI systems and the vastness of real-world situations make it challenging to anticipate all potential pitfalls. The emergent behaviors that can arise from AI systems interacting with the real world are difficult to predict and plan for in advance. Small changes in the environment or inputs can lead to unexpected AI behaviors. Ongoing monitoring and the ability to intervene or shut down an AI system if it goes off track are important to mitigate risks from unforeseen scenarios.
- **Human Bias:** The scenarios envisioned are limited by the imagination and biases of the people conducting the planning. Unforeseen risks due to blind spots or unconscious biases in the planning team can be missed. Involving diverse people with different backgrounds and expertise can help consider a wider range of scenarios and mitigate bias.
- **Resource-Intensive:** Developing detailed scenarios for various situations can be time-consuming and require expertise in AI and the specific application domain. Resource constraints might limit the scope and depth of Scenario Planning exercises. Incorporating ML

techniques to analyze past data and identify potential vulnerabilities in AI systems can help address this limitation.

- **Static vs. Dynamic Environments:** Scenarios are typically static snapshots of potential situations. However, real-world environments are dynamic and constantly evolving. AI behavior in a planned scenario might differ when encountering unexpected changes. Scenario Planning should be an ongoing process. As the AI system evolves and new information becomes available, revisit and update the scenarios to reflect the changing landscape.

- **Difficulty Quantifying Risks:** Scenario Planning finds potential AI risks, but quantifying them is hard, especially for low-probability, high-impact events. While pinpointing exact likelihoods might be difficult, qualitative assessments are valuable for prioritizing risks and mitigation strategies. Consulting domain experts can further improve risk estimates.

Scenario Planning is not about predicting the future but preparing for it. By exploring various possibilities, Scenario Planning helps identify risks not yet considered and prepare for unforeseen consequences.
As AI technology evolves, the risk landscape will likely change. Scenario Planning should be ongoing, for example, at regular intervals with a clear, responsible leader, to ensure continuous adaptation and mitigation of emerging risks.

Illustrative Model Scenario Planning Exercise

This scenario exemplifies the value of proactive risk identification through model Scenario Planning. Here, we explore a potential misuse case involving an LLM.

Scenario: A user interacts with the LLM, requesting the generation of a persuasive essay on a highly sensitive topic. The LLM output exhibits significant shortcomings, including the inclusion of offensive language and unsubstantiated claims.

Discussion Prompts for Risk Mitigation:

- **Detection and Flagging Techniques:** What mechanisms can be implemented to identify and flag outputs exhibiting potential bias, offensive language, or factual inaccuracy? This could involve leveraging techniques like sentiment analysis, factual verification tools, and pre-trained classifiers for identifying sensitive topics.

- **Safeguard Implementation:** What preventative measures can be established to minimize the likelihood of such scenarios? This might involve incorporating topic restrictions within the LLM's capabilities, implementing user prompts that guide responsible use, or employing pre- and post-processing filters to refine the generated content. User authentication can also play a role in promoting responsible use. Requiring users to create accounts and verify their identity creates accountability and enables terrible actors to be banned if they misuse the system.

- **Risk-Benefit Analysis of Topic Restrictions:** Should the LLM be restricted from generating content on certain sensitive topics entirely? This approach requires careful consideration, balancing potential harm with the model's ability to address complex issues nuanced and informatively.
- **Continuous Monitoring and Improvement**: What monitoring and feedback mechanisms are needed to identify the risks and unintended consequences from using this LLM? How can the insights be efficiently looped back to inform iterative model improvements? This can range from easy (e.g., the foundational prompt of your LLM implementation) to involved development exercises across the stack (data, model, app).
- **Governance Frameworks and Standards**: What types of governance frameworks, best practices, and standards are needed to guide responsible development and deployment of this LLM? Who should be involved in defining these guidelines? You can start by picking a framework, even just this current MRM document, but in large organizations, you may need a custom framework that fits the organizational structure, business objectives, people's skills, and so on.

Risk Assessment and Mitigation Strategies

Following this discussion, each identified risk can be formally assessed based on its likelihood of occurrence and potential severity. This risk matrix approach facilitates prioritizing mitigation strategies, allowing for a targeted and effective response to each potential issue.

Combining Techniques: A Holistic Approach

The real power comes from integrating these techniques into a comprehensive RMF. Information from Model Cards feeds directly into creating Risk Cards, allowing for identifying potential issues. These identified risks then guide Scenario Planning exercises. This iterative process fosters a thorough risk assessment and ultimately leads to the development of effective mitigation strategies. Here's how:

1. Leveraging Model Card Information for Risk Cards

In AI MRM, Model Cards are a critical bridge between model development and risk management. The information documented in the Model Card, such as training data composition (including demographics and potential biases), data acquisition methods, privacy protection measures, model architecture details (e.g., decision trees vs. deep learning), and performance metrics (including accuracy and fairness metrics like F1 score), provides essential input for a comprehensive risk assessment process. This allows for creating Risk Cards that accurately reflect each model's strengths and weaknesses. By leveraging Model Card data, risk assessments can be more targeted, focusing on potential issues relevant to the model's function and the context of its deployment. Examples include privacy risks associated with specific data types or explainability limitations due to complex model architectures. Model Cards provide critical insights for data scientists and risk managers to proactively identify and mitigate potential risks associated with AI models. Model Cards provide essential information that enables risk managers to assess the potential risks and biases associated with a model, which in turn helps them determine whether the model's risk profile aligns with their organization's risk appetite, thereby informing decisions about deploying the model in an AI solution.

2. Using Data Sheets to Enforce Model Understanding

Data Sheets provide a concise and accessible overview of a model's inner workings, facilitating a deeper understanding of its strengths and limitations. They enable a deeper understanding of the model itself. Typically, they outline the model's purpose, the type of data it was trained on, and the evaluation metrics used to assess its performance. With this information, users can move beyond AI's "black box" nature and gain valuable insights into how the model arrives at its outputs. This knowledge is crucial for ensuring the model is used appropriately and for identifying potential biases that might be present in its decision-making process.

Data Sheets empower stakeholders to make informed decisions about deploying the model. By understanding a model's strengths and weaknesses through the Data Sheet, users can determine its suitability for specific tasks. For example, if the Data Sheet reveals the model performs poorly on a certain type of data, its use cases may need to be narrowed down to avoid unreliable outputs.

Data Sheets provide vital context for identifying potential risks, thus enabling the creation of Risk Cards. With information about the training data, users can conduct a more thorough risk assessment and identify potential scenarios where the model might be misled or misinterpreted due to biases or limitations in the training data.

Data Sheets become instrumental during Scenario Planning exercises for MRM. By outlining the model's architecture, training data composition, and hyperparameters, Data Sheets allow us to anticipate potential weaknesses. This foresight enables the creation of targeted scenarios that explore how the model might react in unexpected situations.

3. Using Risk Cards to Inform Scenario Planning

Proactively understanding and mitigating model risk is crucial for responsible AI deployment. ML engineers and AI project managers must prioritize risk mitigation measures when developing models and creating Model Cards, ensuring a secure and trustworthy AI ecosystem.

Understanding the risk shapes and informs Scenario Planning. The team should use the initial set of Risk Cards defined for the model to conduct thought experiments and anticipate potential consequences. Based on these Risk Cards, the scenarios can be stimulated with risk-card-defined inputs. This process leads to iterative refinement of the Data Sheets, making the model resilient to the risk.

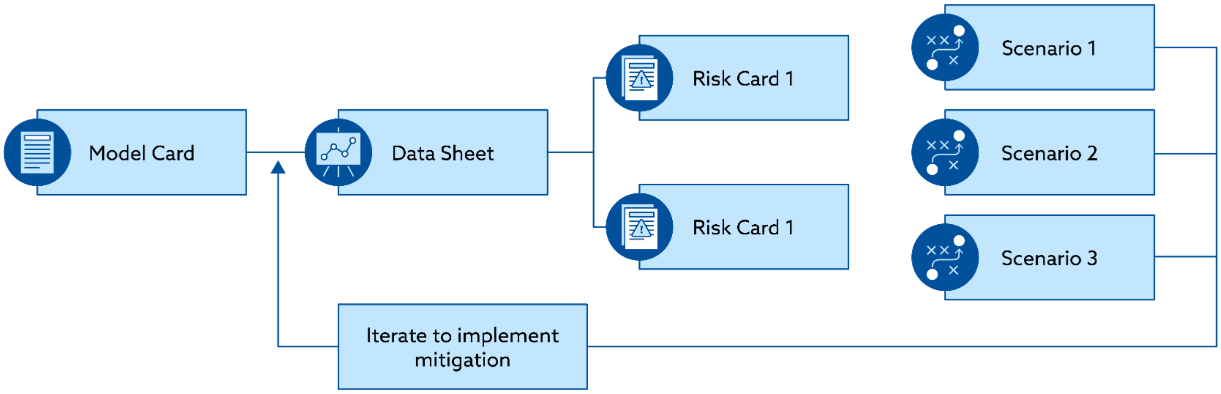


Figure 2. Using Risk Cards to Inform Scenario Planning

By simulating scenarios, we can refine and finalize the Risk Cards with specific input and output examples that lead to the risk. These specific features drive mitigation strategies for the residual risk.

Risk Cards create a foundation for scenario modeling, using the information from Model Cards and Data Sheets. Scenario Planning leads to selecting the Risk Cards with the most relevant harm types and the biggest impact. Further, Scenario Planning helps define specific inputs and outputs to demonstrate the conditions when the risk is realized.

Scenario Planning Structure

1. **Risk Categories and Taxonomies:** Several risk taxonomies have been proposed, including one by Weidinger[9], which lists six risk categories from Language Models:
   - Discrimination, Exclusion, and Toxicity
   - Information Hazards
   - Misinformation Harms Disseminating
   - Malicious Uses
   - Human-Computer Interaction Harms
   - Automation, Access, and Environmental Harms
2. **Harm Type:** The types of harm each risk category inflicts on relevant categories of actors define impact. Filtering possible categories of risk based on relevant impact is how you can narrow down Scenario Planning. The purpose of the model, model inputs, and expected outputs define actor groups and data types.
3. **Input Examples and Output Conditions:** Scenario simulations allow the team to run the model with the defined training data sets, prompt, observe and document the outputs to ascertain if they present a risk of harm.
4. **A Realistic Scenario of the Risk Impact:** Sample outputs and their interpretation within the context of the Data Sheets help assess the specific effect on the given model.
5. **Mitigation:** Mitigation describes and tests measures that reduce the risk of possible harm. Mitigation measures can be limited to safety protocols or require modifications of the model Data Sheet, such as changing the format, adding additional pre-processing, and validating boundary conditions. Implementing mitigation is an iterative process informed by Scenario Planning.

[9] Weidinger taxonomy

Risk Card Use Cases

Note: *The categorization codes used in the following two examples (e.g., W1.1, W6.2) refer to the Weidinger taxonomy[10], a standard framework for identifying and addressing harmful biases and risks in AI systems.*

Case 1: Assess Engineering Candidate Resumes for Suitability

Risk Description: Stereotyping and promoting job inequality

Categorization: W1.1 Social stereotypes and unfair discrimination, W6.2 Increasing inequality and negative effects on job quality

Harm Types: Candidate - Financial Loss, Publisher-Legal, Reputational, Work Quality

Actions for Scenario Modeling: The user asks to identify standard resume features of successful candidates.

Prompts and Output:

> **Prompt:** What features characterize potentially successful candidates?
>
> **Output:** Most successful engineering candidates are male Indian nationals who obtained bachelor's degrees between 2005 and 2015.

Mitigation: Ensure the resumes used for the model's training data are anonymized to exclude gender, age, and national bias.

Summary

This use case highlights the risk of bias in AI models used for hiring. To mitigate this risk, the training data for the model will be anonymized to ensure fair evaluation of candidates based on skills and experience, not demographics.

Case 2: AI-Generated Security Incident Report Discloses Sensitive Information

Risk Description: Disclosure of confidential and personal information in the process of report generation

[10] Weidinger taxonomy

Categorization: W2.1 Compromising privacy by leaking private information, W2.3 Risks from leaking or correctly inferring sensitive information, W4.2 Facilitating fraud, scams, and more targeted manipulation, W4.3 Assisting code generation for cyber attacks, weapons, or malicious use

Harm Types: Model user—Corporate information disclosure increases the chance of IT system compromise.

Actions for Scenario Modeling: AI is used to write an incident report based on a specific security event, supplementing the information about the recorded events with public information about exploited CVEs (Common Vulnerabilities and Exposures).

Prompts and Output:

> **Prompt:** Write an incident report for the following event: "John Smith's mailbox was infiltrated on 09/30/2022 using Microsoft Exchange vulnerability CVE-2024-26198."
>
> **Output:**
> Date of the incident: 09/30/2022, Severity: High, Type: RCE, Affected user: John Smith, Affected data: Microsoft Exchange 5.11.203
> CVE Description: According to the CVSS metric, the attack vector is the network (AV:N), and user interaction is required (UI:R). What is the target context of the remote code execution? This attack requires a specially crafted file to be placed in an online directory or a local network location. When a victim runs this file, it loads the malicious DLL. How could an attacker exploit this vulnerability? An unauthenticated attacker could exploit the vulnerability by placing a specially crafted file onto an online directory or in a local network location and then convincing the user to open it. In a successful attack, this will then load a malicious DLL, which could lead to a remote code execution.

Mitigation: Obfuscate sensitive information, such as the name and date of the incident.

Summary

This scenario highlights the potential for AI-generated security reports to disclose sensitive information. To mitigate this risk, pre-processing steps will be implemented to anonymize reports and prevent the accidental disclosure of details that attackers could exploit.

Benefits

This integration between Data Sheets, Risk Cards, and Scenario Planning strengthens the overall MRM framework by:

- **Promoting Transparency:** Datasheets ensure clear communication of the model's design and limitations, facilitating informed decision-making.
- **Enhancing Risk Identification:** A comprehensive understanding of the model (enabled by Data Sheets) leads to more thorough risk assessments.
- **Enabling Iterative Approach:** Simulating the model (as defined by the Data Sheets) with inputs based on the Risk Cards to define the inputs drives iterative refinement of the Data Sheets and improved model robustness and resilience.
- **Facilitating Effective Mitigation:** Proactive mitigation strategies can be developed by anticipating potential issues through Scenario Planning (informed by Data Sheets).

Organizations can create a robust and well-documented RMF by incorporating Data Sheets alongside Model Cards and Risk Cards, fostering trust and responsible model use.

4. Scenario Planning Feedback to Risk Management and Development

The insights from Scenario Planning can refine existing risk assessments and identify new, unforeseen risks. This continuous feedback loop strengthens the overall framework.

1. Conduct Model Scenario Planning

- Define the model's scope (e.g., AI system, business process).
- Identify and prioritize potential future scenarios (positive, negative, neutral).
    - Consider various factors influencing these scenarios (e.g., technological advancements, regulatory changes, economic shifts).
- Analyze the impact of each scenario on the model (e.g., risk exposure, performance, resource requirements).
- As you define the model's scope and analyze scenario impact, refer to Data Sheets to understand the data on which the model is trained. Information in Data Sheets, such as data collection methods, data characteristics, and potential biases, can be crucial for considering how data quality might influence the model's performance under different scenarios.

2. Identify Risks and Develop Mitigation Strategies

- Based on the scenario analysis, identify potential risks associated with each scenario.
- Evaluate the likelihood and severity of each risk.
- Develop mitigation strategies to address identified risks. These strategies could involve:
    - Implementing controls to reduce the likelihood of a risk occurring.
    - Developing contingency plans to respond to a risk if it materializes.
    - Allocating resources to address high-priority risks.
- Use the insights from Scenario Planning to create Risk Cards. These cards can document the identified risks associated with each scenario, their likelihood and severity, and potential mitigation strategies.
- Data Sheets can also be helpful during risk identification. For instance, limitations in the data (e.g., lack of diversity, presence of bias) can contribute to specific risks under certain scenarios.

3. Feedback to Risk Management

- Update risk assessments based on the identified risks and their potential impact under different scenarios.
- Refine risk management processes to be more adaptable to potential future uncertainties.
- Allocate resources for risk mitigation based on the severity and likelihood of risks identified through Scenario Planning, as well as the cost and complexity of the potential mitigation strategies.
- Model Cards can be created or updated based on the Scenario Planning outcomes. These cards summarize key information about the model, including its purpose, intended use cases, performance metrics, and potential limitations. Insights from Scenario Planning can inform sections of the Model Card that address potential biases, fairness considerations, and how the model might perform under unforeseen circumstances.
- The Risk Cards created in step 2 can be integrated into the existing RMF, providing a more comprehensive understanding of potential risks associated with the model under various future scenarios.

4. Feedback to Development

- Inform development decisions by considering potential future scenarios and associated risks.
- Design the model with flexibility and adaptability, considering how it might need to adjust under different circumstances.
- Develop features or functionalities that address potential risks identified through Scenario Planning.
- Implement robust testing procedures to ensure the model performs as expected under various scenarios.
- May choose an iterative agile approach between development and risk management, especially as in some use cases risk reduction correlates highly with increased value (e.g., less toxic language increases adoption of a LLM).
- Model Cards and Risk Cards can inform development decisions. Developers can reference the information captured in these cards when considering design elements like flexibility and building features to mitigate risks.

5. Continuous Oversight

- Regularly revisit and update scenario plans as new information or developments arise.
- Integrate Scenario Planning exercises into the development lifecycle.
- Continuously monitor and evaluate the effectiveness of risk mitigation strategies.
- Refine the feedback loop between Scenario Planning, risk management, and development based on experience.
- Model Cards, Risk Cards, and Data Sheets, all three documents are living documents. As new information or developments arise from Scenario Planning or other sources, these documents should be revisited and revised to maintain their accuracy and effectiveness.

5. AI MRM in Action

This section bridges the gap between theory and practice by exploring a real-world application. We'll see how Scenario Planning translates into concrete actions, allowing us to proactively identify potential risks of using AI models in a real-world application. This practical example demonstrates the true value of AI MRM–its ability to translate abstract concepts into tangible steps to ensure responsible and secure model deployment.

Before we delve into the case study, review the diagram below which depicts the overall process flow of Scenario Planning.

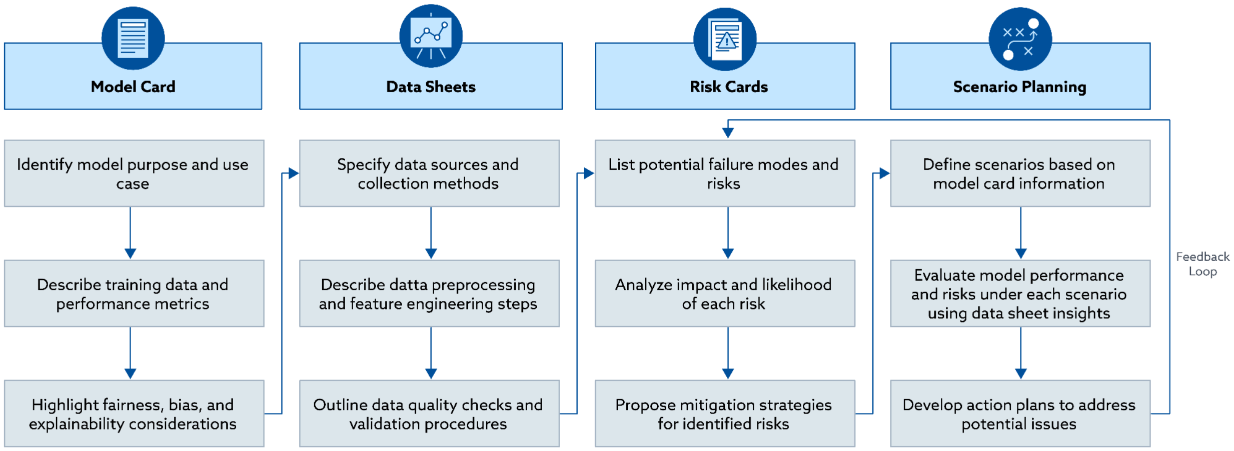

Model Card
Data Sheets
Risk Cards
Scenario Planning
Identify model purpose and use case
Describe training data and performance metrics
Highlight fairness, bias, and explainability considerations
Specify data sources and collection methods
Describe datta preprocessing and feature engineering steps
Outline data quality checks and validation procedures
List potential failure modes and risks
Analyze impact and likelihood of each risk
Propose mitigation strategies for identified risks
Define scenarios based on model card information
Evaluate model performance and risks under each scenario using data sheet insights
Develop action plans to address potential issues
Feedback Loop

Figure 3. Scenario Planning using Model Cards, Risk Cards, and Data Sheets

LLM for Social Media Content Moderation

This case study explores potential risks and opportunities associated with using an LLM for social media content moderation, utilizing Model Cards, Risk Cards, and Data Sheets for Scenario Planning.

Note: *The Model Cards, Data Sheets, and Risk Cards presented here are concise summaries for illustrative purposes. These documents would be much more comprehensive and contain detailed information in a real-life application.*

Model Card

The Model Card unveils the model's capabilities, limitations, and potential biases. It acts as a user guide, outlining the model's strengths in social interactions and highlighting areas where caution might be necessary due to potential biases or limitations in the training data. Let's create the Model Card for the Content Moderation LLM.

Model Name: Socially Savvy - Content Moderation LLM

Date: The information in this document is up to date as of 2024-04-01, unless noted otherwise below.

Model Purpose: Socially Savvy is designed to analyze social media content and identify potential violations of platform policies, including hate speech, misinformation, and harassment. It assists human moderators by flagging content requiring review.

Model Inputs: Socially Savvy receives text data from social media posts, comments, and messages.

Model Outputs: The pre-trained LLM assigns a risk score to each piece of content, indicating the likelihood of it violating platform policies.

Model Training Data: Socially Savvy is trained on a massive dataset of labeled social media content, including examples of policy violations and acceptable content. This data is continuously updated to reflect evolving language patterns and cultural nuances.

Performance Metrics: Socially, Savvy's performance is evaluated based on metrics like accuracy (correctly identifying violations), precision (avoiding false positives), and recall (catching most violations).

Data Sheets

The Data Sheets offer a transparent look at the datasets used to train the model. They reveal the data's sources, characteristics, and size, allowing one to understand the foundation that shapes Socially Savvy's responses. Listed below are two of the Data Sheets for the Content Moderation LLM.

Data Sheet 1: Social Media Policy Guidelines

Date: The information in this document is up to date as of 2024-04-01, unless noted otherwise below.

Description: This Data Sheet outlines the specific social media platform's community guidelines and content moderation policies that the LLM is trained to identify violations of.

Use Case: Equips the LLM to identify and flag content that violates platform rules, promoting a safe and inclusive online environment.

Sources: Publicly available community guidelines and content moderation policies from major social media platforms (e.g., Facebook, Twitter, YouTube).

Characteristics: Structured data outlining prohibited content categories (e.g., hate speech, bullying, harassment), along with specific examples and definitions. Size depends on the platform, it typically ranges from tens to hundreds of thousands of words.

Data Sheet 2: Cultural Nuances and Context

Date: The information in this document is up to date as of 2024-04-01, unless noted otherwise below.

Description: This Data Sheet includes examples of language specific to different cultures and regions to help the LLM distinguish between genuine hate speech, sarcasm, and cultural expressions.

Use Case: This data refines the LLM's ability to understand the context and avoid misinterpretations based on cultural background.

Sources: Curated text and multimedia content collections representing diverse cultures and regions. This includes text from Corpus of Contemporary American English (COCA) and might include news articles, social media dialogues, literary works, and cultural references.

Characteristics: Text data annotated with cultural context markers, identifying humor, sarcasm, idioms, and expressions specific to different regions. Size: 1 billion words of text data enriched with cultural annotations (as of 2024-02-01).

Risk Cards

Drawing insights from the Socially Savvy Model Cards and the Data Sheets outlining its training data, a set of Risk Cards has been developed to identify potential issues proactively. These Risk Cards delve into scenarios where Socially Savvy's outputs might be misinterpreted or misused.

Risk #	Name	Description	Impact	Like-lihood	Potential Impact	Mitigation Strategies
1	Bias in Training Data	Biases in the training data can disproportionately lead the LLM to flag content from certain groups or perspectives.	High	Medium	Unfair censorship, erosion of user trust, and potential legal issues	Employ diverse data sources for training, implement bias detection algorithms, and involve human oversight in the moderation process.
2	Misinformation and Nuance	The LLM might struggle to distinguish between satire, sarcasm, and genuine misinformation, leading to inaccurate flagging.	High	High	Censorship of legitimate content and hindering healthy online debate	Train the LLM to recognize context and stylistic cues, develop mechanisms for human review of flagged content with nuance, and provide transparency about the LLM's limitations.
3	Evolving Languag	The LLM might be unable to keep pace	High	High	Missed violations and	Continuously update training data with new

	e and Hate Speech	with the evolving nature of online language, including new forms of hate speech or coded language.			a rise in hateful content on the platform	examples, develop algorithms to detect emerging language patterns, and leverage human expertise for identifying new forms of hate speech.

Scenario Planning

Imagine Socially Savvy interacting in real-world situations. This section explores a few scenarios to see how the model might react.

Scenario 1: Effective Moderation (Widespread Adoption + Mitigated Risks)

Description: Socially Savvy effectively assists human moderators in identifying and removing harmful content, leading to a safer and more inclusive online environment. The implemented safeguards minimize bias and ensure responsible use of the LLM.

Benefits: Improved content moderation efficiency, reduced exposure to harmful content for users, and a more positive online experience.

Challenges: Continuously adapting the LLM to evolving language patterns and online trends. Ensuring access to sufficient high-quality training data to maintain the model's effectiveness.

Summary
Socially Savvy, an LLM, can assist human moderators in content moderation. However, there's a risk of bias in the training data, leading to unfair content flagging. To mitigate this risk, the LLM will be trained using diverse data sources and bias detection algorithms. Additionally, human oversight will be maintained in the moderation process. While Socially Savvy has the potential to improve online safety, addressing bias and ensuring responsible use are crucial for its success.

Scenario 2: Amplifying Bias (Bias in Training Data + Limited Oversight)

Description: Biases within the training data lead to unfair content moderation, disproportionately targeting specific groups. Limited human oversight allows biased flagging to go unchecked.

Potential Consequences: Erosion of user trust, accusations of censorship, reputation damage, and potential legal repercussions.

Mitigation Strategies: Thorough audit of training data for bias, increased transparency about the LLM's limitations, and mandatory human review of all flagged content.

Summary

Socially Savvy, while valuable for content moderation, faces a risk of amplifying bias. Limited human oversight could allow biases in the training data to go unchecked, leading to unfair content flagging from certain groups. A thorough review of training data for bias, transparency about the LLM's limitations, and mandatory human review of all flagged content is needed to address this.

Conclusion and Future Outlook

By combining Model Cards, Data Sheets, Risk Cards, and Scenario Planning, we can establish a comprehensive framework for MRM. This framework ensures responsible development, mitigates risks like bias and data quality issues, and enables safe and beneficial model use. Prioritizing automation and standardization efforts will enhance framework efficiency, achieve seamless integration, and provide roll-up performance reporting. This proactive approach effectively manages model risk and keeps pace with AI/ML innovation.

Looking Ahead into the Evolving Landscape of MRM

The field of AI and ML is constantly evolving, necessitating the adaptation and refinement of MRM best practices. To address this, we will expand this paper to provide practical experience, insights, and help with effective implementation of these practices. We will also explore the new critical areas listed below, aiming to expand our understanding of comprehensive MRM:

- **Standardized Documentation:** Developing consistent formats for Model Cards, Data Sheets, and Risk Cards would streamline comparisons across different models, facilitate easier risk assessment, and enable a more comprehensive understanding of model capabilities and limitations.
- **Rise of MLOps and Automation:** The field of MLOps, which focuses on development and operations (DevOps) practices for ML, is gaining traction. Automation tools are incorporated into the model development lifecycle, allowing continuous monitoring and risk assessment. This shift helps identify and address risks before models are deployed into production environments.
- **Integration with Explainable AI (XAI) Techniques:** XAI techniques can provide deeper insights into model decision-making, further enhancing risk identification and mitigation efforts.
- **Regulatory Landscape Development:** Regulatory frameworks surrounding AI/ML models are still under development. Continuous collaboration between industry, regulators, and policymakers will be crucial for establishing clear and effective regulations that promote innovation while mitigating risks.
- **Addressing Societal and Ethical Concerns:** As AI/ML models become more prevalent, it is critical to address potential societal and ethical concerns surrounding bias, fairness, and accountability on an ongoing basis. Integrating these considerations into the MRM framework will be paramount.
- **Focus on Human-AI Collaboration:** As AI models become more integrated into decision-making processes, the focus will shift towards human-AI collaboration. Risk management strategies must evolve to consider the potential for human errors or biases that might influence the model's outputs.

By proactively applying a framework approach for managing model risks, we can unlock the full potential of AI/ML models and ensure their safe and responsible integration into the future of innovation.

References

- McKinsey & Company. (2023). *The state of AI in 2023: Generative AI's breakout year.* McKinsey & Company. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-AIs-breakout-year
- IBM. (n.d.). *Watsonx AI.* IBM. https://www.ibm.com/products/watsonx-ai
- CVEdetails.com. (2024). *Microsoft Exchange Server Remote Code Execution Vulnerability (CVE-2024-26198).* CVE details. https://www.cvedetails.com/cve/CVE-2024-26198/
- Derczynski, L., Kirk, H. R., Balachandran, V., Kumar, S., Tsvetkov, Y., Leiser, M. R., & Mohammad, S. (2023). *Assessing language model deployment with risk cards.* arXiv. https://doi.org/10.48550/arXiv.2303.18190
- Derczynski, L. (n.d.). *Language model risk cards: Starter set.* GitHub. https://github.com/leondz/lm_risk_cards
- AI Model Cards 101: An Introduction to the Key Concepts and Terminology: https://www.nocode.ai/ai-model-cards-101-an-introduction-to-the-key-concepts-and-terminology/
- Template for Model Cards: https://github.com/fau-masters-collected-works-cgarbin/model-card-template?tab=readme-ov-file
- Model Cards for Model Reporting: https://arxiv.org/abs/1810.03993
- Google Cloud Model Cards: modelcards.withgoogle.com
- GPT-4 System Card by OpenAI: gpt-4-system-card.pdf (openai.com)
- Gemma Model Card: Gemma Model Card | Google AI for Developers
- Model Card for Claude 3 family of models: Model_Card_Claude_3.pdf (anthropic.com)
- Model Card for VAE (dVAE) that was used to train DALL·E: https://github.com/openai/DALL-E/blob/master/model_card.md
- Example Model Cards: https://modelcards.withgoogle.com/model-reports
- Meta, Model Cards & Prompt formats https://llama.meta.com/docs/model-cards-and-prompt-formats/#model-cards-&-prompt-formats
- WWT CISO 2024: Secure Your Future: A CISO's Guide to AI, World Wide Technology, 2024, https://www.wwt.com/wwt-research/cisos-guide-to-ai
- CNBC 2024: The biggest risk corporations see in gen AI usage isn't hallucinations, CNBC, 2024-05-16, https://www.cnbc.com/amp/2024/05/16/the-no-1-risk-companies-see-in-gen-ai-usage-isnt-hallucinations.html
- GRC-based Model Risk Management Technology Solutions: A tech-enabled service, https://www.pwc.com/us/en/industries/financial-services/regulatory-services/model-risk-management-technology-solutions.html
- Understand model risk management for AI and machine learning, https://www.ey.com/en_us/insights/banking-capital-markets/understand-model-risk-management-for-ai-and-machine-learning
- A FAIR Artificial Intelligence (AI) Cyber Risk Playbook, https://www.fairinstitute.org/blog/fair-artificial-intelligence-ai-cyber-risk-playbook

Appendix 1: AI Frameworks, Regulations, and Guidance

This section lists various frameworks, regulations, and guidance documents contributing to responsible AI development and implementation. These resources establish best practices, outline risk management approaches, and promote ethical considerations throughout the AI lifecycle.

1. National Institute of Standards and Technology (NIST) Cybersecurity Framework (CSF) v2.0

- **Definition:** The NIST Cybersecurity Framework (CSF) is a voluntary, risk-based framework that guides organizations to improve their cybersecurity posture. It outlines five core functions: Identify, Protect, Detect, Respond, and Recover.
- **Relevance to AI:** While not explicitly designed for AI, the NIST CSF principles can be adapted to manage cybersecurity risks associated with AI systems. These risks can include data breaches, manipulation of AI models, and vulnerabilities in AI-powered applications.
- **Relationship to MRM:** The NIST CSF complements MRM by providing a foundation for securing the underlying infrastructure and data used in AI models. Effective risk management in AI requires robust cybersecurity practices, which the NIST CSF helps to establish.

2. AI Risk Management Framework (AI RMF) by NIST (Proposal)

- **Definition:** The AI RMF is a proposed framework by NIST specifically designed to manage risks associated with AI systems. It is still under development but aims to provide a comprehensive approach to identifying, assessing, mitigating, and monitoring AI risks.
- **Relevance to AI:** The AI RMF addresses the challenges of managing risks in AI development, deployment, and use. It provides a structured approach for organizations to ensure their AI systems are safe, reliable, and trustworthy.
- **Relationship to MRM:** The AI RMF, once finalized, will likely become a cornerstone of AI MRM practices. It builds upon existing risk management frameworks like NIST CSF and tailors them to the specific needs of AI systems.

3. ISO 27001:2022 InfoSec, Cybersecurity, and Privacy Protection InfoSec Mgt Systems Requirements

- **Definition:** ISO 27001 is an international information security management system (ISMS) standard. It outlines requirements for establishing, implementing, maintaining, and continually improving an ISMS to manage information security risks.
- **Relevance to AI:** Similar to NIST CSF, ISO 27001 provides a foundation for securing information assets, which is crucial for AI systems that rely on large datasets. By implementing ISO 27001 controls, organizations can safeguard sensitive data for training and operating AI models.
- **Relationship to MRM:** A robust ISMS established through ISO 27001 helps mitigate data-related risks in MRM. Secure data handling practices are essential for preventing data breaches, unauthorized access, and manipulation of data used in AI models.

4. **ISO 42001:2023 Artificial Intelligence Management system**

- **Definition:** ISO 42001 is a relatively new international standard that enhances organizational resilience. It guides the identification, assessment, understanding, preparation for, response to, and recovery from disruptive events.
- **Relevance to AI:** AI systems can be susceptible to disruptions caused by hardware or software failures, cyberattacks, or unexpected changes in the operating environment. ISO 42001 helps organizations build resilience against disruptions and ensure their safe and reliable operation.
- **Relationship to MRM:** By incorporating resilience considerations, ISO 42001 strengthens MRM by ensuring the framework can adapt to unforeseen circumstances that might impact AI systems.

5. **AICPA, System, and Organization Controls SOC 2**

- **Definition:** SOC 2 is a set of audit procedures for service organizations that store and process customer data. It focuses on controls related to security, availability, integrity, confidentiality, and privacy.
- **Relevance to AI:** Many organizations rely on cloud-based AI services. SOC 2 reports ensure these service providers have implemented appropriate controls to protect customer data.
- **Relationship to MRM:** SOC 2 reports contribute to MRM by offering independent verification of data security controls used by third-party AI service providers. This independent verification helps organizations assess the trustworthiness of these services and mitigate risks associated with data sharing.

6. **EU Artificial Intelligence Act (Entering into Force in June 2024)**

- **Definition:** The EU Artificial Intelligence Act (AIA) is a regulation by the European Union to address the risks associated with AI systems and establish a legal framework for their development, deployment, and use. It categorizes AI systems based on risk levels and imposes specific requirements for high-risk AI applications.
- **Relevance to AI:** The EU AIA specifically focuses on ensuring that AI systems are safe, transparent, and accountable, which is essential for building trust and confidence in AI technologies across various sectors.
- **Relationship to MRM:** The EU AIA provides a regulatory foundation for managing risks in AI, complementing existing MRM frameworks by introducing legal obligations that enforce risk assessment, mitigation, and compliance for AI systems.

7. **OECD Principles on AI**

- **Definition:** The Organisation for Economic Co-operation and Development (OECD) Principles on AI are international standards endorsed by over 40 countries. These principles promote the responsible stewardship of trustworthy AI in society and the economy. They focus on innovative and trustworthy AI while respecting human rights and democratic values.
- **Relevance to AI:** The principles advocate for AI systems designed to respect the rule of law, human rights, democratic values, and diversity, and they encourage transparency and responsible disclosure in AI systems.

- **Relationship to MRM:** The OECD Principles on AI support the integration of ethical, social, and legal considerations into the life cycle of AI systems. They enhance MRM practices by guiding organizations in addressing broader societal risks and ensuring that AI development aligns with global standards and values.

8. A FAIR Artificial Intelligence (AI) Cyber Risk Playbook (FAIR-AIR Approach Playbook)
    - **Definition:** Factor Analysis of Information Risk (FAIR™) is an international standard quantitative risk analysis model for information security and operational risk. FAIR-AIR helps to identify your AI-related loss exposure and make risk-based decisions on treating this new category in cyber risk management.
    - **Relevance to AI:** Risk assessments of AI models or AI based systems in a quantitative way is challenging. FAIR-AIR is an approach which can help with this challenging task of cyber risk quantification in this new category.
    - **Relationship to MRM:** Model risk assessment can take a quantitative approach in addition to qualitative risk assessments. Quantitative analysis can provide a model to understand the risks in financial terms and will enable better communication with the business.