OFFICIAL DOCUMENT
V1.0: 09/30/2024

PREPARED BY BABL AI INC.
& THE ALGORITHMIC BIAS LAB

EU AI Act: Quality Management System Oversight Certification

2024 Handbook

babl

Table of Contents

Introduction

Quality management system oversight is a critical component of ensuring compliance with the EU AI Act, which governs the development and deployment of artificial intelligence (AI) systems. As AI becomes more prevalent and integrated into various aspects of our lives, it is important to ensure that these systems adhere to the strict regulatory requirements set forth by the EU AI Act, promoting transparency, fairness, and accountability.

Managing a quality management system for AI involves establishing, maintaining, and improving processes to ensure that AI systems meet the necessary regulatory standards. By overseeing these systems, quality managers can provide assurance to stakeholders that AI systems are operating as intended, support compliance with the EU AI Act, and further BABL's mission of promoting human flourishing in the age of AI.

To fulfill this mission, AI quality managers need a number of skills and capabilities, many of which fall outside the standard quality management skillset. Current quality management standards are critically important for maintaining regulatory compliance and professional conduct. However, further capabilities are needed to deal with the complex sociotechnical systems in which modern AI and machine learning algorithms are embedded. This is due to a) the rapid advance of the underlying technology, b) our rapidly evolving understanding of the ways these systems can affect society (both negatively and positively), and c) the fact that sufficient knowledge of (a) and (b) is needed to mitigate risk when managing quality systems for AI.

What this training is

The objective of this training is to equip quality managers with sufficient knowledge to identify relevant risks of using AI systems, best practices for the governance of such systems, and the common techniques and workflows that are involved in modern AI/ML development. This knowledge will be used to ensure compliance with the EU AI Act while evaluating quality management documentation, under the supervision of a more experienced quality manager, and is meant to lay the foundation for further on-the-job learning opportunities.

What this training is not

This training is not meant to encompass all the knowledge and competencies needed to manage a quality management system for AI in the absence of supervision or oversight from a more experienced risk or compliance professional working at a firm with a mature system of quality management.

Program Courses

The courses in the program center around the four core competencies that are critical to performing effective management of a quality management system for AI. The subject matter associated with some of these courses is dynamic and rapidly changing, so we've focused where possible on larger conceptual and operational frameworks that will allow quality managers to absorb and retain new knowledge in these areas. These frameworks, coupled with on-the-job training, are enough for entry-level AI quality managers to work on teams, and can also form the foundation for many positions in responsible AI, AI governance, and the risk management of complex AI and algorithmic systems.

The core courses are, in the preferred order in which they should be taken (though this is only a suggestion):

EU AI Act Conformity Requirements for High Risk AI Systems
Algorithmic Risk & Impact Assessments
AI Governance & Risk Management
Bias, Accuracy, & the Statistics of AI Testing

Each of these courses consists of lectures and exercises that take approximately two weeks of dedicated time to finish.

Courses

Below we describe the subject matter of the core courses in the Certificate Program, outline the competencies expected after completion of the course, and link to resources to help prepare and review for the certification exam.

EU AI Act Conformity Requirements for High Risk AI Systems

In the EU AI Act Conformity Requirements for High-Risk AI Systems Course, you will gain an understanding of the EU AI Act, focusing on the identification of high-risk AI systems, obligations for developers, strategies for implementing requirements, and achieving conformity through assessments. The goal of this course is not to present a full legal reading of the Act, but rather to prepare risk, legal, and compliance professionals to develop and oversee critical compliance efforts required for high-risk AI systems. The focus will be on practical implementation strategies.

Competencies

1. Knowing the language of the Act and be able to navigate the text
2. Understanding the risk classification of the EU AI Act and performing basic risk categorization of AI systems, being able to identify high risk AI Systems
3. Gaining a basic understanding of the Conformity Requirements for High Risk Systems
4. Understanding the basic strategies for compliance including Quality Management System, Transparency & Information Provision, Risk Management and Conformity Assessments.
5. Being able to evaluate transparency statements to assess whether they are compliant with the EU AI Act
6. Understanding different elements of AI testing, validation, and monitoring
7. Being able to apply some basic strategies for compliance and development of internal controls

Resources

Exercise1 (required):
Exercise #1: Risk Categorization | The Algorithmic Bias Lab
Exercise2 (required):
https://courses.babl.ai/courses/eu-ai-act-conformity-requirements-for-high-risk-ai-systems/lectures/53674368
Regulation - EU - 2024/1689 - EN - EUR-Lex

The Act Texts | EU Artificial Intelligence Act
EU AI Act Compliance Matrix
Ethics Guidelines for Trustworthy AI | FUTURIUM | European Commission
Risk Management in the Artificial Intelligence Act | European Journal of Risk Regulation | Cambridge Core
https://babl.ai/wp-content/uploads/2024/04/BABL_ISO_42001_Primer.pdf

Algorithmic Risk & Impact Assessments

This course teaches you a systematic approach to algorithmic risk and ethical impact assessments, which is a necessary skill for practitioners in the space of emerging technology. The primary focus of this course is on BABL AI's particular framework for detecting and evaluating algorithmic risk and impact, and connections to other impact assessment frameworks are made in the AI Governance & Risk Management course.

Competencies

1. Identify the socio-technical components of an algorithmic system that are relevant for risk analysis
    a. Contextual use-case factors, human in/on/over the loop, UI/UX, user training, transparency considerations, etc.
2. Produce a narrative of these components (a "CIDA" narrative) as a form of algorithmic transparency
3. Identify important stakeholders
4. List engagement strategies for relevant stakeholders to determine their salient interests, and rights, and identify potential harms due to the algorithmic system
    a. Understand the methods and limitations for prioritizing harm or interests
5. Decide which components of the algorithmic system can serve as metrics for risk analysis (related to 1 above)
6. Recognize common technical metrics (overlap with Bias, Accuracy, & the Statistics of AI Testing course below)
7. Develop initial assessment strategies for these metrics

Resources

BABL AI Risk & Impact Assessments I
BABL AI Risk & Impact Assessments II
BABL AI Risk Assessment Cheat Sheet
ADS Questionnaire
Risk Assessment Requirements Question List

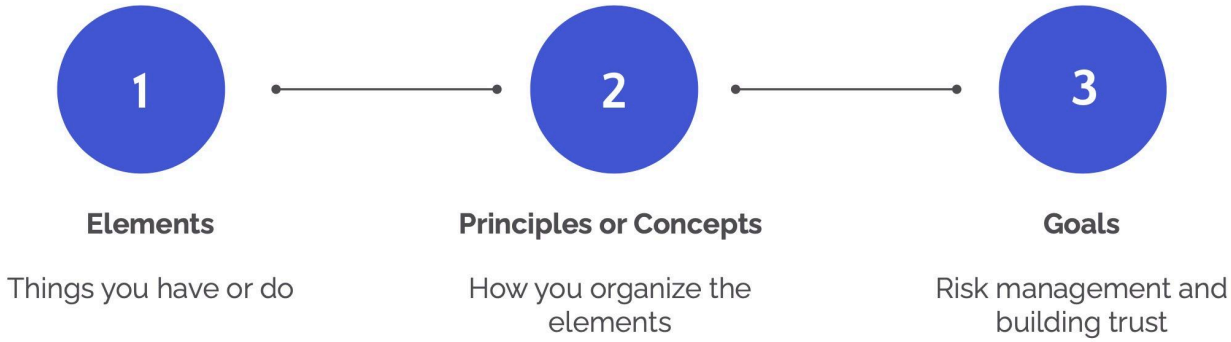
Risk Assessment Spreadsheet Template
Exercise #1 (required)
Exercise #2 (required)

AI Governance & Risk Management

This course covers topics in AI governance, risk management, and emerging regulations of algorithmic systems needed for AI auditors. Given the vast array of risk management frameworks emerging and the rapidly changing nature of the regulatory environment, this course focuses on first principles for governance and risk management, using current examples to illustrate the kinds of controls an auditor may encounter in different organizations.

Competencies

1. Know the purpose of risk management
    a. "Enhance an organization's ability to achieve its mission, vision, strategic objectives, and strengthen its competitive position by **limiting downside risk**, **ensuring quality** of the organization's products and services, and **building trust** with stakeholders critical to the success of the mission"
2. Understand how the building blocks of an organization's resources come together to form an AI governance and risk management system
    a. Elements, principles/concepts, and goals (see figure below)
    b. Elements: combinations of people, activities, and assets
3. List popular risk management frameworks most relevant for AI systems
    a. NIST AI Risk Management Framework
    b. COSO ERM
    c. SR11-7 Model Risk Management
    d. ISO 31000, 23894, 42001
    e. ForHumanity AI RMF
4. Recognize the major emerging laws and regulations and their high-level requirements
    a. EU – AI Act, DSA, GDPR
    b. US – Federal Regulator effort (EEOC, FTC, etc.), State/Local (NYC Local Law 144, etc.), Federal efforts (AI Bill of Rights, Executive orders, Algorithmic Accountability Act, etc.)



Resources

BABL AI - The Current State of AI Governance
BABL AI - Quick Guide to AI Governance
NIST AI RMF
ForHumanity AI RMF
ISO 42001
SR 11-7 Model Risk Management
Exercise #1 (required)
Exercise #2 (required)

Bias, Accuracy & the Statistics of AI Testing

This course focuses on the technical testing of AI and machine learning systems and covers topics such as bias testing/mitigation, validation, and basic principles behind performance metrics, uncertainty, and statistics. Topics are covered at a depth sufficient to allow non-technical auditors to communicate with technical specialists, and highlight important critical perspectives that technical specialists need to focus on in order to maintain professional skepticism and explore the full range of algorithmic parameters relevant to risk.

Competencies

1. Understand the importance of conducting a risk assessment to inform technical testing
2. Know basic concepts of statistics relevant to technical testing
    a. Distributions and sampling
    b. Central value estimation (mean, median, etc.)
    c. Uncertainty estimation (standard deviation, standard error, etc.)
    d. Basic hypothesis testing (t-test, non-parametric methods)
3. Analyze testing methodologies using "parameter-space thinking"
    a. Comparing test coverage with use-case-specific input parameter space

4. Know basic performance metrics and how they are constructed
5. Understand the purpose and examples of robustness testing, and methods for improvement
    a. Adversarial and stress testing
    b. Techniques for improvement: regularization, adversarial training, ensemble methods
6. Understand the concepts of validity and validation testing, e.g.,
    a. Statistical vs. substantive validity
    b. Internal vs. external validity

Resources

BABL AI - Bias Testing Cheat Sheet
Basic Statistics
Technical Testing Notes
Machine Learning for High-Risk Applications
Exercise #1 (required)
Exercise #2 (required)

Credentials

The primary purpose of this certificate program is to equip students with the base-level skills required to begin working as AI Quality Managers. However, many of the competencies presented above are foundational to responsible AI and risk management in general and would thus be useful for career advancement separately from managing quality management systems. We are therefore issuing Certificates of Completion for each of the core cores separately, a Professional Certificate in AI Quality Management for completion of all four courses, and (under limited circumstances) certifications for certified professional AI Quality Managers.

Certificates of completion

These will be given for each of the core courses, and requires students to:

1. Watch all the lectures
2. Complete all quizzes at with a 70% or higher score
3. Complete all exercises and projects (usually two per course) at a level sufficient to demonstrate understanding of the material (at the discretion of the course instructor)

For those that complete all four of the core courses they will receive a Professional Certificate in AI Quality Management. As it takes a significant amount of proficiency and work to complete all the course assignments at a satisfactory level, this credential represents a demonstration of many of the capabilities necessary for several entry-level Quality Management and Responsible AI positions.

Certified AI Quality Manager Certification

Under some circumstances, we will allow students to take a certification exam to become a Certified AI Quality Manager. The primary purpose of this course is to ensure that BABL AI team members have the required knowledge and skills to work on ai quality management, but other students may be eligible to obtain a certification if:

1. They have obtained an AI Quality Manager Professional Certificate;
2. They sign on to adhere to the ForHumanity Certified Auditor Code of Ethics;

The certification will need to be renewed each year through a short refresher exam and/or continuing education.

Certification Exam

The certification exam consists of case studies with associated questions, both multiple choice and short answer, with approximately 40 questions total. The exam will be offered approximately once per month and will be proctored over Zoom. While the majority of the content will be non-technical, there will be some technical questions from Bias, Accuracy & the Statistics of AI Testing course. Any necessary equations will be given in the question text.

Example Questions

Example Case Study
Applicant-Role Matching Algorithm in HR Using Embedding Vectors

Context:
Human Resources (HR) departments often grapple with the high volume of applicants for a variety of job roles. The manual screening process is not just time-consuming but also prone to inconsistencies and human biases. This case study focuses on an algorithm aimed at automating the initial stages of the recruitment process.

The algorithm compares resumes and job descriptions to rank candidates for specific roles. This technology is highly applicable for businesses and HR firms dealing with a large number of applications across diverse positions.

Architecture
The core of the algorithm involves transforming text-based information from job descriptions and resumes into a form that can be compared mathematically.

Data and Development Process
The model relies on a pre-trained word embedding model, which has been trained on a large text corpus. The algorithm takes the pool of resumes and job descriptions, processes them through the text-to-vector transformation stage (an "encoder"), and then calculates similarity scores.

Testing
The success of the algorithm is gauged by human experts who review the rankings for a sample of job roles. They assess whether the top-ranked candidates seem appropriate for the positions based on their resumes. After trying several word embedding models, the experts found a model that matched how they would rank the candidates.

Question 01: After testing for bias, the company reports for this job-candidate matching algorithm, the selection rate for male candidates is 40%, and the selection rate for female candidates is 30%. What is the Disparate Impact Ratio for female candidates based on these selection rates?

- ☐ A) 30/(40+30) = 0.43
- ☐ B) 30/40 = 0.75
- ☐ C) 40/30 = 1.33
- ☐ D) 40 /(40+30) = 0.57

Question 02: Assume that the testing from Question 01 was done prior to deploying the system, so the company does not have access to actual hiring outcomes, just the raw rankings. What assumption did the company need to make in order to calculate these selection rates? Choose the best answer.

- ☐ A) That the ranking algorithm is entirely free from any form of bias
- ☐ B) That the top-ranked candidates above some cutoff are always the ones who get selected
- ☐ C) That all candidates, regardless of gender, have the same level of qualification
- ☐ D) That the algorithm's rankings are purely based on skill sets and not influenced by external factors like networking

Question 3: The "core" of NIST's AI Risk Management Framework is broken down into four "functions", which are:

- ☐ A) Develop, Detect, Mitigate, Maintain
- ☐ B) Govern, Map, Measure, Manage
- ☐ C) Plan, Execute, Deploy, Monitor
- ☐ D) Govern, Detect, Mitigate, Monitor

Question 4: Under the EU AI Act, what conditions must be met for an AI system to be classified as high-risk?

- ☐ A. The AI system is designed to entertain users and is sold independently.
- ☐ B. The AI system is a safety component of a product or a regulated product under EU legislation and must undergo a third-party assessment.
- ☐ C. The AI system is open-source and available for public use.
- ☐ D. The AI system is used for administrative tasks within an organization.

Question 5: Which of the following AI systems would not be classified as high-risk under the EU AI Act?

- ☐ A. An AI system designed to significantly influence human decision-making processes.
- ☐ B. An AI system that profiles individuals.
- ☐ C. An AI system designed for specific, limited tasks and does not influence human judgment.
- ☐ D. An AI system that serves as a safety component for a regulated product.

Question 6: What must providers do if they believe their AI system listed in Annex III is not high-risk?

- ☐ A. They must discontinue the system immediately.
- ☐ B. They must document and register their assessment, making it available to national authorities upon request.
- ☐ C. They must apply for an exemption from the AI Board.
- ☐ D. They do not need to take any action as long as the system is safe.

Question 7: Who is responsible for providing guidelines and a detailed list of what constitutes high-risk and non-high-risk AI applications?

- ☐ A. The European Parliament.
- ☐ B. The AI Board.
- ☐ C. The Commission, after consulting with the AI Board.
- ☐ D. National authorities in each EU member state.

Question 8: How can the Commission update the criteria for determining high-risk AI systems?

- ☐ A. By issuing a public referendum.
- ☐ B. Through delegated acts if new evidence suggests changes are needed, without reducing protection levels.
- ☐ C. By consulting with technology companies only.
- ☐ D. By creating a new set of rules every five years.

Professional Code of Conduct

We adhere to several key references when shaping our professional Code of Conduct. Our goal is to eventually implement the Code of Conduct from the International Association of Algorithmic Auditors (IAAA), which is currently under development.

For now, we are using standards from the ForHumanity Code of Conduct and the International Code of Ethics for Professional Accountants. These resources guide our commitment to ethical practices, integrity, and professionalism in the field.

ForHumanity Certified Auditor Code of Ethics;

Glossary of Terms

EU AI Act Conformity Requirements for High Risk AI Systems

AI system

Machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.

Conformity assessment

The process of demonstrating whether the requirements set out in Chapter III, Section 2 relating to a high-risk AI system have been fulfilled.

Deployer

A natural or legal person, public authority, agency or other body using an AI system under its authority except where the AI system is used in the course of a personal non-professional activity.

Fundamental Rights Impact Assessment (FRIA)

The aim of the fundamental rights impact assessment is for the deployer to identify the specific risks to the rights of individuals or groups of individuals likely to be affected, identify measures to be taken in the case of a materialisation of those risks. The impact assessment should be performed prior to deploying the high-risk AI system, and should be updated when the deployer considers that any of the relevant factors have changed.

General purpose AI model

An AI model, including when trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable to competently perform a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications. This does not cover AI models that are used before release on the market for research, development and prototyping activities.

High-Risk AI Systems

AI Systems which are intended as safety components of products or are themselves products and undergoing third-party conformity assessment; AI Systems used in predefined areas specified in the regulation (e.g., biometric identification, management of critical infrastructure, education, employment, essential private and public services, law enforcement, etc.).

Human Oversight

High-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which they are in use. Human oversight shall aim to prevent or minimise the risks to health, safety or fundamental rights that may emerge when a high-risk AI system is used in accordance with its intended purpose or under conditions of reasonably foreseeable misuse.

Limited Risk AI Systems

"Limited Risk" AI Systems don't really exist as a category in the EU AI Act, but people usually use it for AI systems that have transparency requirements but are not high risk.

Minimal Risk AI System

All other AI systems that do not fall under the prohibited, high-risk or limited risk categories. allowed to be developed and used according to the discretion of the provider, as long as they comply with the Act's provisions.

Operator

A provider, product manufacturer, deployer, authorised representative, importer or distributor.

Post-market monitoring system

All activities carried out by providers of AI systems to collect and review experience gained from the use of AI systems they place on the market or put into service for the purpose of identifying any need to immediately apply any necessary corrective or preventive actions.

Prohibited AI Practices

AI Systems deploying subliminal techniques beyond a person's consciousness; exploiting vulnerabilities due to age, physical or mental disability, or social status; creating or expanding facial recognition databases through untargeted scraping; providing social scoring of individuals; predictive policing.

Provider

A natural or legal person, public authority, agency or other body that develops an AI system or a general-purpose AI model or that has an AI system or a general-purpose AI model developed and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge.

Quality Management System

Providers of high-risk AI systems shall put a quality management system in place that ensures compliance with this Regulation. That system shall be documented in a systematic and orderly manner in the form of written policies, procedures and instructions.

Risk Management System

A risk management system shall be established, implemented, documented and maintained in relation to high-risk AI systems. The risk management system shall consist of a continuous iterative process run throughout the entire lifecycle of a high-risk AI system, requiring regular systematic updating.

Safety component

A component of a product or of an AI system which fulfils a safety function for that product or AI system, or the failure or malfunctioning of which endangers the health and safety of persons or property.

Algorithmic Risk Assessment & Management

Algorithm Auditing:
Algorithm Auditing is the process of evaluating and scrutinizing algorithms, particularly those used in Automated Decision Systems, to ensure fairness, accountability, transparency, and the absence of bias. Audits assess the data, design, implementation, and impact of algorithms, aiming to identify and mitigate unintended consequences, discriminatory practices, and other ethical issues. Algorithm Auditing fosters trust, compliance with regulations, and the responsible development and deployment of AI technologies.

Bias Testing:
Bias Testing in AI involves evaluating models for any unfair or unequal treatment of different groups based on sensitive characteristics like race, gender, or age. By analyzing model predictions and outcomes across diverse groups, practitioners can identify and address biases, ensuring that models do not perpetuate or amplify existing inequalities and stereotypes.

Fairness Metrics:
Fairness Metrics are quantitative measures used to evaluate the fairness of AI models across different groups or individuals. They help in identifying disparities in model outcomes, such as differences in error rates, prediction thresholds, or benefits between groups. Common fairness metrics include demographic parity, equalized odds, and disparate impact.

False Positive Rate:
False Positive Rate is a performance metric that quantifies the proportion of negative instances incorrectly classified as positive by a classification model. It is essential for assessing the model's reliability, especially in applications where false positives have significant consequences, such as medical diagnosis, fraud detection, and criminal justice.

False Negative Rate:
False Negative Rate measures the proportion of positive instances that a classification model incorrectly identifies as negative. Managing this rate is crucial in scenarios where failing to identify positive cases can lead to adverse outcomes, such as missing a disease diagnosis or failing to identify a security threat.

Statistical Distribution:
A statistical distribution is a representation of the frequencies or probabilities of potential outcomes in a set of data. It can be visualized using graphs or histograms to depict how often each outcome occurs. Common distributions include the normal distribution, binomial distribution, and Poisson distribution, each describing different types of data patterns and variability.

Mean, Median, and Central Values:
The mean is the average of a set of numbers, calculated by adding all values and dividing by the count of values. The median is the middle value in a sorted list of numbers, separating the higher half from the lower half. Both are measures of central tendency, representing central values in a distribution, but they might differ in skewed distributions.

Standard Deviation:
Standard deviation is a measure of the amount of variation or dispersion in a set of values. A low standard deviation indicates that the values tend to be close to the mean, while a high standard deviation indicates that the values are spread out over a wider range. It is a key metric in statistics for quantifying uncertainty and variability.

Statistical Significance:
Statistical significance is a measure used to assess whether the results of an experiment or study are unlikely to have occurred by chance. Typically, a result is deemed statistically significant if the p-value (probability value) is less than a predetermined threshold, commonly 0.05, indicating that the observed effect is likely not due to random variation.

Hypothesis Testing:
Hypothesis testing is a statistical method used to make inferences about population parameters based on a sample of data. It involves formulating a null hypothesis (assuming no effect or no difference) and an alternative hypothesis, calculating a test statistic from the sample data, and determining the probability (p-value) of observing such a statistic under the null hypothesis. If the p-value is below a predetermined threshold, the null hypothesis is rejected in favor of the alternative hypothesis, suggesting that the observed effect is statistically significant.

Use-Case-Relevant Data Collection:
Use-Case-Relevant Data Collection emphasizes the importance of gathering data that is pertinent and representative of the specific task or problem the AI model is designed to address. This involves considering the diversity, quality, and volume of the data, ensuring that it covers various scenarios, demographics, and conditions. This approach minimizes biases and inaccuracies, resulting in models that are more robust, fair, and generalizable to real-world applications.

Algorithmic Risk/Impact Assessment:
Algorithmic Risk/Impact Assessment is a comprehensive evaluation of the potential risks and impacts associated with deploying AI models and algorithms. It includes assessing biases, ethical considerations, data privacy, security vulnerabilities, and the societal and

individual implications of algorithmic decisions, aiming to mitigate unintended consequences and ensure responsible AI use.

Stakeholder Interests:
Stakeholder Interests refer to the diverse needs, concerns, and expectations of individuals or groups who have a stake or interest in a particular technology, project, or organization. In the context of AI and machine learning, stakeholders can include developers, users, regulators, affected communities, and the general public. Balancing stakeholder interests is crucial for the ethical and responsible development and deployment of AI technologies, ensuring that the systems are fair, transparent, beneficial, and aligned with societal values and norms. Addressing stakeholder interests involves engaging in dialogue, incorporating diverse perspectives, assessing potential impacts, and mitigating risks associated with AI applications.

Risk Likelihood and Magnitude Analysis:
Risk Likelihood and Magnitude Analysis involves evaluating the probability of various risks associated with AI applications and estimating their potential impact. This analysis helps in prioritizing risks and developing appropriate mitigation strategies, ensuring the responsible and ethical deployment of AI technologies.

AI Risk Management Framework:
An AI Risk Management Framework provides structured guidelines and practices for identifying, assessing, mitigating, and monitoring risks associated with the development and deployment of AI systems. It encompasses considerations related to data quality, model robustness, ethical use, compliance with regulations, and societal impact, aiming to foster responsible and trustworthy AI. See NIST AI RMF for an example.

Model Risk Management:
A discipline focused on managing risks arising from the development, deployment, and use of quantitative models, including AI and machine learning models, especially in the financial services sector. It involves validating model assumptions, testing for biases and inaccuracies, monitoring model performance over time, and ensuring that models are used appropriately within their intended applications. See SR 11-7.

Transparency:
Transparency in AI refers to the openness and clarity regarding the development, deployment, and decision-making processes of AI models. It involves providing understandable and accessible information about model architecture, training data, algorithms, and objectives, enabling stakeholders to scrutinize and trust the AI system.

Explainability:
Explainability refers to the ability of an AI model to provide understandable and interpretable reasons for its predictions and decisions. Explainable models help build trust, facilitate model debugging and improvement, and ensure that users and stakeholders can understand and challenge model outcomes, promoting accountability and ethical use of AI.

Interpretability:
Interpretability is the degree to which a human can comprehend and make sense of the decisions made by an AI model, or the degree to which one can predict the model's output given a new input. An interpretable model allows users to assess the model's reliability and validity, fostering trust and enabling more informed and responsible use of AI technologies.