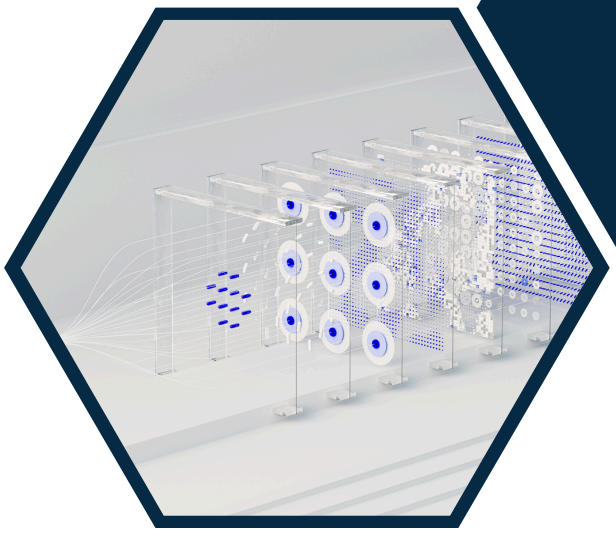
babl

OFFICIAL DOCUMENT
V1.0: 10/22/2024

PREPARED BY BABL AI INC.
& THE ALGORITHMIC BIAS LAB

AI & Algorithm Auditor Certificate Program



2024 Handbook

Table of Contents

Introduction

Algorithm auditing is a critical component of ensuring the safe and ethical development and deployment of artificial intelligence (AI) systems. As AI becomes more prevalent and integrated into various aspects of our lives, like in the rapid adoption of generative AI and large language models like ChatGPT, it is important to ensure that these systems are transparent, fair, and accountable.

Algorithm auditing involves examining algorithmic systems and their governance to ensure that ethical, safety, and compliance risks are sufficiently managed. By conducting independent audits, algorithm auditors can provide assurance to stakeholders that these systems are operating as intended, promote transparency and accountability, and further BABL's mission of promoting human flourishing in the age of AI.

In order to fulfill this mission, algorithm auditors need a number of skills and capabilities, many of which fall outside the standard audit/assurance skillset. Current financial audit and assurance standards are critically important for maintaining independence 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 evaluating algorithmic systems.

What this training is

The objective of this training is to equip auditors with sufficient knowledge to identify relevant risks of using algorithmic 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 assess the risk of material misstatement while evaluating audit documentation, under the supervision of a more experienced auditor, and is meant to lay the foundation for further on-the-job learning opportunities (either at BABL AI or other algorithm auditing firms).

What this training is not

This training is not meant to encompass all knowledge and competencies needed to perform algorithm auditing in the absence of supervision or oversight from a more experienced auditor working at a firm with a mature system of quality management.

Program Courses

The courses in the program center around the five core competencies that are critical to performing effective algorithm audits. 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 auditors to absorb and retain new knowledge in these areas. These frameworks, coupled with on-the-job training, are enough for entry-level auditors to work on audit 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):

Algorithmic Risk & Impact Assessments
AI Governance & Risk Management
Algorithms, AI, & Machine Learning
Bias, Accuracy, & the Statistics of AI Testing
Algorithm Auditing & Assurance

Each of these courses consists of lectures, reading material and quizzes that take approximately one week 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.

Algorithms, AI, & Machine Learning

This is a technical crash course in Automated Decision (Augmentation) Systems and generative AI with a focus on bringing non-technical consultants, risk, and policy professionals up to speed on these emerging technologies. The goal is to gain a sufficient understanding of modern techniques to perform risk analysis, governance, and basic algorithm auditing.

Competencies

1. List and understand the most common techniques used in AI and machine learning
    a. Recognize the difference between supervised, unsupervised, and reinforcement learning approaches to machine learning
    b. Understand the difference between a model and the training algorithm that created the model in machine learning
    c. Recognize the variety of ways in which AI/ML applications can be categorized, e.g.,
        i. Deterministic vs. stochastic
        ii. Domain use (hiring algorithms, social media recommender systems, facial analysis use cases, medical diagnoses, etc.)
        iii. Input type (computer vision, natural language processing, feature vectors, multi-modal)
        iv. Output type (categorical classification, regression, text generation, image generation, actions, multi-modal outputs)
2. Understand the methods, data, and resources needed to create machine learning and statistical models for automated decision systems (ADMs) and generative AI.
    a. Recognize common model architectures used in machine learning
        i. Deep neural networks, convolutional neural networks (CNN), recurrent neural networks (RNN, LSTM, etc.), transformers, etc.
    b. Understand the importance of gradient descent and loss functions
    c. Identify common architecture components and their functions
        i. Linear layers, activation functions, encoders, decoders, attention layers
    d. Know the different functions of training data, validation data, and test data
    e. Know the common language of large language models and chatbots

        i. Prompts, context windows, supervised fine-tuning, reinforcement learning with human feedback, embeddings, vector databases
    f. List common Python libraries that are typically used by machine learning engineers, e.g.,
        i. TensorFlow, PyTorch, sci-kit learn, pandas
3. Identify select technical value judgments that must be made in the development of AI/ML
    a. Use-case-relevant data collection
    b. Model selection, validation
4. Recognize elements of basic Python code

Resources

https://courses.babl.ai/courses/1510825/lectures/34642336

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

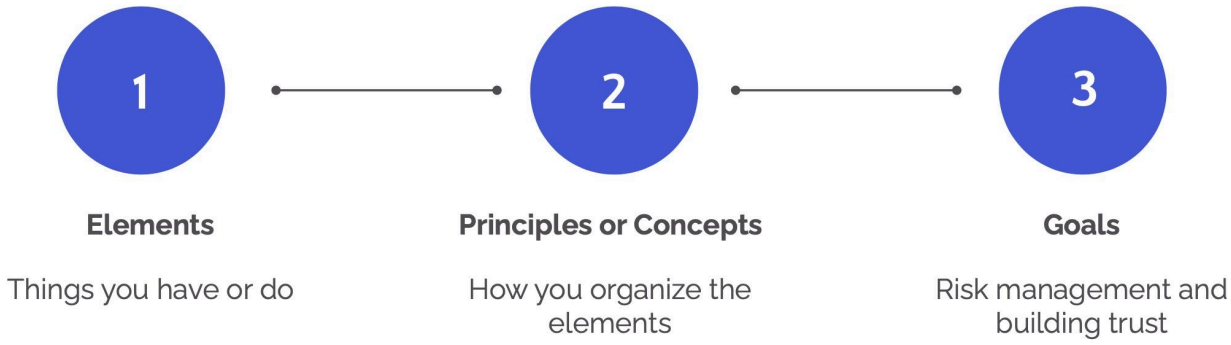
AI Governance & Risk Management

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

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

Algorithm Audit & Assurance

This course focuses on the actual practice of algorithm audit and assurance. The field of regulatory assurance has a long history, and most of the standards from that field can and should be applied in the context of algorithmic systems. However, the complexity and rapidly evolving nature of modern algorithmic systems that utilize AI/ML present new challenges for audit and assurance practitioners; this course focuses on how BABL navigates these challenges.

Competencies

1. Understand the difference between audit, assurance, and assessment
2. Know the goals of an audit/assurance engagement
3. Know the basics of independence requirements and where to find more information
4. Understand the relationship between the subject matter, relevant criteria and type of engagement

   a. Recognize the difference between limited and reasonable assurance engagements
   b. Recognize the difference between attestation and direct engagements in the context of who is evaluating the subject matter against the relevant criteria
   c. List the requirements for suitable criteria: relevance, completeness, reliability, neutrality, and understandability
5. Understand that audit planning requires assessing the risk of material misstatement
   a. Recognize that for algorithmic systems, factors that influence the risk include:
      i. The complexity of the system
      ii. The level of internal controls that the company employs
6. List the types of audit procedures and evidence
   a. Recognize the difference between inspection, observation, confirmation, re-calculation, re-performance, analytical procedures, and inquiry
   b. Recognize what sufficiency and appropriateness mean in terms of audit evidence, and that appropriateness captures the concepts of relevance and reliability
7. Know the types of opinions in an assurance engagement
   a. Recognize the difference between a reasonable assurance opinion and a limited assurance opinion

Resources

ISAE 3000 (Revised)
Taxonomy of AI Audit & Assurance & Assessment
For Humanity FHCA Code of Ethics
Sarbanes Oxley Act of 2002
IASSB - ISQM
PCAOB - AS 1105
BABL AI comments on draft rules DSA
Handbook of the International Code of Ethics for Professional Accountants

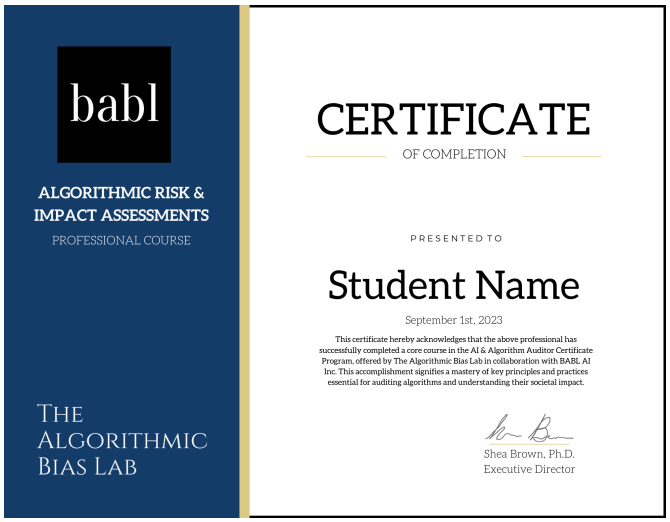
Credentials

The primary purpose of this certificate program is to equip students with the base-level skills required to begin working as AI and algorithm auditors. 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 algorithm auditing. We are therefore issuing Certificates of Completion for each of the core cores separately, a Professional Certificate in AI & Algorithm Auditing for completion of all five courses, and (under limited circumstances) certifications for professional auditors.

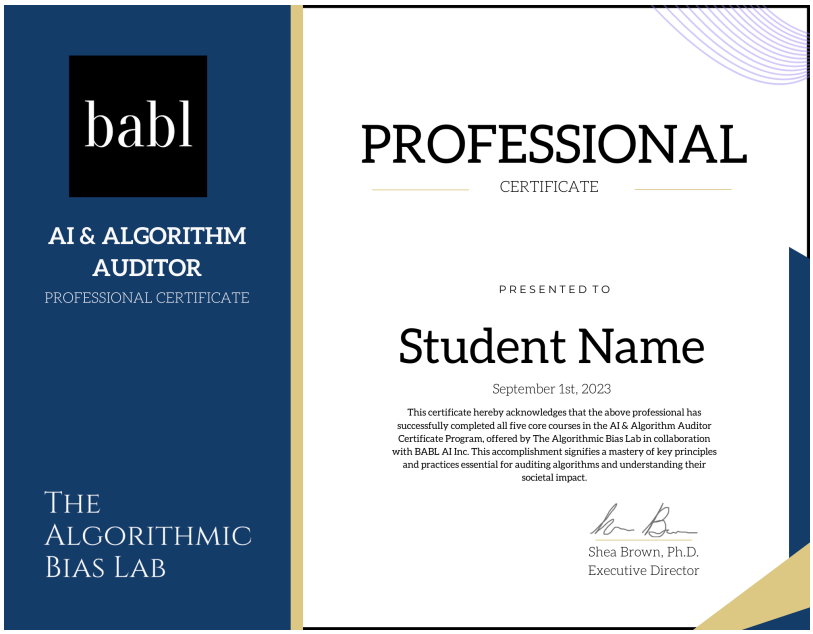
Certificates of completion

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



For those that complete all five of the core courses they will receive a Professional Certificate in AI & Algorithm Auditing. As it takes a significant amount of proficiency and work to complete all the courses at a satisfactory level, this credential represents a demonstration of many of the capabilities necessary for several entry-level Responsible AI positions. As such, the recommended professional titles are either **AI & Algorithmic Risk Specialist** or **AI & Algorithm Audit Specialist.**



Professional Auditor Certification

Under some circumstances we will allow students to take a certification exam to become a Certified AI & Algorithm Auditor. The primary purpose of this course is to ensure that BABL AI auditors have the required knowledge and skills to work on our audit teams, but other students may be eligible to obtain a certification if:

1. They have obtained an AI & Algorithm Auditor Professional Certificate;
2. They sign on to adhere to the ForHumanity Certified Auditor Code of Ethics;
3. They have completed the Capstone Project at a satisfactory level (70% or higher score)

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

Capstone Project

The Capstone Project for the AI & Algorithm Auditor Certification Program is a comprehensive, hands-on course that integrates key exercises from all five core courses. Designed for those seeking full certification, this capstone will challenge participants to apply their knowledge across diverse areas, including AI technologies, risk assessment, governance frameworks, bias testing, and audit practices. By completing this project,

participants will demonstrate their ability to assess, evaluate, and ensure compliance and ethical standards for AI systems, solidifying their skills as certified AI & Algorithm Auditors.

Capstone course page

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 the Algorithms, AI & Machine Learning and Bias, Accuracy & the Statistics of AI Testing courses. Any necessary equations will be given in the question text.

Example Case Study and Questions

Machine Learning 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. To do this, the algorithm utilizes word embeddings to convert the textual information into high-dimensional vectors. For this use case, they are utilizing those provided by OpenAI's `text-embedding-ada-002` embedding. After this transformation, it calculates a similarity score between each job description and resumes to rank applicants.

1. Pre-processing: Resumes and job descriptions undergo tokenization and cleaning to remove non-essential information.

2. Word Embeddings: Textual elements are transformed into vectors using the pre-trained model. These vectors capture the semantic essence of the words in the text.

4. Similarity Measurement: A mathematical technique is employed to calculate the similarity between the aggregated vectors representing the job descriptions and resumes. This score serves as a basis for ranking applicants for each role. For this use case they utilized a cosine similarity score.

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: Does the Applicant-Role Matching Algorithm described above use supervised, unsupervised, or reinforcement learning methods to determine the ranking?
A) Supervised learning
B) Unsupervised learning
C) Reinforcement learning

Question 02: Word embeddings are known to sometimes capture biases present in the corpus of text they were trained on. In what way are these biases typically manifested?
A) They only capture statistical properties, thus are entirely neutral
B) They favor technical jargon words over simpler non-technical words
C) They can capture societal biases like gender or racial stereotypes
D) They prioritize the meaning of nouns over verbs

Question 03 (short answer): Describe one way that the algorithm and/or methodology described above could lead to the risk of discrimination. Your answer must identify a) which group could be at risk; b) what feature of the algorithm or development process is causing that risk; and c) the exact mechanism that could lead to discriminatory rankings.

Question 04 (short answer): For the risk of discrimination that you listed above, describe a method for testing for this discrimination. Your description must include a) the data needed for the measurement and b) a numerical metric that you could use to measure the potential for discrimination.

Question 05: Using a similarity metric to rank candidates based on their resume and the job description makes what conceptual assumption? (Choose the most correct answer)
A) All words used in resumes and job descriptions have universally understood meanings
B) The most similar resume to the job description is always the best fit for the job
C) All candidates have uniformly structured resumes
D) The algorithm will inherently remove all biases in the recruitment process

Question 06: 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 07: Assume that the testing from Question 06 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 08: What is the primary objective of obtaining sufficient appropriate evidence in an assurance engagement under ISAE 3000?
A) To make the engagement more time-consuming
B) To reduce the practitioner's risk to zero
C) To support the practitioner's assurance conclusion
D) To comply with regulatory requirements

Question 09: When performing a limited assurance engagement under ISAE 3000, what kind of procedures are generally performed?
A) More limited than for a reasonable assurance engagement, primarily consisting of inquiries and analytical procedures.
B) Exactly the same procedures as for a reasonable assurance engagement.
C) Solely external confirmations.
D) Comprehensive testing of transactions, data, and disclosures.

Question 10: 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

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

Algorithms, AI, & Machine Learning

Algorithms:
Algorithms are step-by-step computational procedures or formulas used to solve problems and perform tasks in computer science. They serve as the foundation for all computational processes, including those in artificial intelligence and machine learning. In machine learning, algorithms, such as decision trees, clustering, and neural networks, are employed to learn patterns from data and make predictions or decisions without being explicitly programmed for the task.

Artificial Intelligence:
Artificial Intelligence (AI) refers to the simulation of human intelligence in machines. It encompasses a broad range of technologies and approaches that enable machines to perceive, reason, learn, and make decisions, aiming to replicate or mimic human cognitive functions. AI can be categorized as narrow or weak AI, which is designed for a specific task, and general or strong AI, which has the potential to perform any intellectual task that a human being can.

Machine Learning:
Machine Learning (ML), a subset of AI, is a field of study focused on developing algorithms and models that enable computers to learn from data and improve their performance over time without being explicitly programmed. ML techniques are used to uncover patterns, relationships, and insights from large volumes of data, which can be used to make predictions, classify objects, recognize speech, translate languages, and many other tasks. ML encompasses various subfields, including supervised learning, unsupervised learning, and reinforcement learning, each with its methods and applications.

Automated Decision Systems (ADS):
Automated Decision Systems (ADS) leverage algorithms, data, and models to autonomously make decisions or predictions without human intervention. These systems are used across various sectors, including finance, healthcare, and criminal justice, to optimize processes, allocate resources, and assess risks. While ADS can enhance efficiency and objectivity, they also raise concerns regarding fairness, accountability, and transparency, necessitating rigorous evaluation and regulation.

Supervised Learning:
Supervised Machine Learning (ML) is a type of ML where the algorithm is trained on a labeled dataset, which means that each training example is paired with an output label. The

algorithm receives a set of inputs along with the corresponding correct outputs, and it learns by comparing its actual output with the correct outputs to find errors and modify the model accordingly. Supervised learning is used for various tasks such as classification, regression, and anomaly detection, with applications in image recognition, speech recognition, and predictive analytics.

Unsupervised Learning:
Unsupervised Machine Learning involves modeling with datasets that don't have labeled responses. The system tries to learn the patterns and the structure from the input data without any supervision. This type of ML can uncover previously unknown patterns in data, but it can be more challenging to understand and validate. Common unsupervised learning techniques include clustering, dimensionality reduction, and association rule learning, used for tasks like market basket analysis, customer segmentation, and feature learning.

Semisupervised Learning:
Semisupervised Machine Learning is a middle ground between supervised and unsupervised learning. In semisupervised learning, the algorithm is trained on a dataset that contains both labeled and unlabeled data. Generally, a small amount of data is labeled while a large amount of data is unlabeled. Semisupervised learning is useful when acquiring a fully labeled dataset is expensive or time-consuming, and it can lead to improved model performance by utilizing the additional unlabeled data.

Self-supervised Learning:
Self-supervised learning is an unsupervised learning paradigm where the data itself provides supervision. Unlike supervised learning, it doesn't rely on external labels. Instead, it automatically generates labels from the data, often by defining a pretext task where the model is trained to predict part of the input data from other parts of the input data. For example, a self-supervised model might learn to predict the next word in a sentence, treating the rest of the sentence as context. This approach allows the model to learn rich representations of the data, which can be useful for a variety of downstream tasks. Self-supervised learning has been particularly successful in natural language processing and computer vision, enabling models to leverage large amounts of unlabeled data effectively.

Reinforcement Learning:
Reinforcement Learning (RL) is a type of machine learning where an agent learns to make decisions by interacting with an environment. The agent receives feedback in the form of rewards or penalties, and it aims to maximize the cumulative reward over time by discovering the optimal policy or strategy. RL is particularly suited for sequential decision-making problems where the optimal solution is not apparent and requires exploration, with applications in game playing, robotics, and autonomous navigation.

Generative AI:
Generative AI refers to a subset of AI technologies capable of creating new content, such as images, text, or music, by learning patterns, structures, and features from input data. Examples include Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), which have applications in art creation, data augmentation, and synthetic data generation. Ethical considerations for Generative AI include the potential for creating deepfakes, copyright infringements, and ensuring the responsible use of generated content.

Deterministic vs. Stochastic:
Deterministic algorithms produce the same output given the same input, operating under fixed instructions and conditions. In contrast, stochastic algorithms incorporate randomness, resulting in varying outputs for the same input. In AI and machine learning, deterministic approaches offer predictability and reproducibility, while stochastic methods, like the outputs of most LLMs, are valuable for exploring diverse solutions, modeling uncertainty, and navigating complex, non-deterministic environments.

Model Architecture:
Model architecture refers to the structured arrangement and connectivity of layers, nodes, and operations in a machine learning model. It defines the way input data is processed and transformed through various layers to produce the output. The architecture is integral to the model's ability to learn patterns and make predictions or classifications and includes components like convolutional layers in CNNs, recurrent layers in RNNs, and attention mechanisms in Transformers.

Hyperparameters:
Hyperparameters are external configurations for machine learning models that are not learned from the data but are set prior to the training process. Examples include learning rate, batch size, and the number of hidden layers. Hyperparameter tuning is a critical step in model development, as the optimal set of hyperparameters can significantly impact the model's learning ability and performance.

Convolutional Neural Networks (CNN):
CNNs are a class of deep learning models primarily used for image recognition and computer vision tasks. They employ convolutional layers with sliding filters to extract hierarchical features from input images. Pooling layers are also used for dimensionality reduction and to make the model translation-invariant, allowing CNNs to efficiently learn spatial hierarchies and patterns in visual data.

Recurrent Neural Networks (RNN):

RNNs are neural networks designed for sequential data processing and prediction, widely used in NLP and time series analysis. They maintain a hidden state that can capture information from previous time steps, allowing them to model temporal dependencies. However, traditional RNNs suffer from vanishing gradient problems, limiting their ability to learn long-range dependencies.

Long Short-Term Memory (LSTM):
LSTM networks are a type of RNN specifically designed to address the limitations of vanilla RNNs in learning long-range dependencies. LSTMs use a system of gates to control the flow of information to be remembered or forgotten at each time step, making them effective for tasks like language modeling, machine translation, and speech recognition.

Loss Function:
A loss function quantifies how well a machine learning model's predictions match the true target values. It is a mathematical function that the model tries to minimize during the training process. The choice of the loss function is critical as it influences how the model is optimized and ultimately, its performance on the task.

Linear Layers:
Linear layers, also known as fully connected or dense layers, are foundational components of neural networks. In these layers, each input is connected to each output by a weight, and a bias is typically added. The linear transformation is often followed by a non-linear activation function, enabling the model to learn complex, hierarchical representations.

Encoders and Decoders:
In sequence-to-sequence models, encoders process the input sequence and compress it into a fixed-size representation called context vector. Decoders then use this representation to generate the output sequence. This architecture is widely used in machine translation, text summarization, and image captioning, where input and output sequences can be of different lengths.

Attention Layers:
Attention layers enable models to weigh and prioritize different parts of the input sequence, focusing on relevant information for making predictions. They are a key component of Transformer models, providing a mechanism to capture long-range dependencies and relationships in the data, leading to improved performance on various NLP tasks.

Training Data:
Training data is the dataset used to train a machine learning model. It consists of input-output pairs, and the model learns to map inputs to outputs by minimizing the loss

function. The quality and diversity of the training data significantly influence the model's ability to generalize to unseen data.

Validation Data:
Validation data is a separate subset of the dataset, not used in the training process, employed to evaluate the model's performance during training and hyperparameter tuning. It helps in identifying overfitting and selecting the best model configuration, ensuring that the model generalizes well to unseen data.

Test Data:
Test data is a dataset used to assess the performance of the final machine learning model after training and validation. It should not be used in any part of the model development process to ensure an unbiased evaluation of the model's generalization ability to new, unseen data.

Gradient Descent:
Gradient Descent is an optimization algorithm commonly used in machine learning and deep learning for minimizing the loss function. The algorithm iteratively adjusts the model's parameters in the direction of the steepest decrease in the loss function. In each iteration, the gradient (derivative) of the loss function with respect to the parameters is computed, and the parameters are updated proportionally to the negative of the gradient. The proportionality factor is determined by the learning rate, a hyperparameter that needs to be carefully chosen. Variants of gradient descent include Stochastic Gradient Descent, which updates parameters using a single data point at each iteration, and Mini-Batch Gradient Descent, which uses a small batch of data points for each update. By progressively refining the model parameters, Gradient Descent helps the model learn the underlying patterns in the training data, thus improving its performance.

Performance Metrics:
Performance metrics quantify the effectiveness of a machine learning model in making predictions or classifications. Common metrics include accuracy, precision, recall, F1 score, and area under the ROC curve (AUC-ROC). The choice of metric depends on the task, and it is crucial for evaluating and comparing model performance.

Model Selection and Validation:
Model Selection and Validation are crucial steps in the development of machine learning models, involving choosing the appropriate model architecture and validating its performance. Model selection is based on the nature of the problem, data characteristics, and the desired outcome, while validation assesses the model's accuracy, reliability, and generalization using unseen data, helping prevent overfitting and underfitting.

Support Vector Machines:
Support Vector Machines (SVM) are a class of supervised learning models used for classification and regression. They work by finding the hyperplane that best separates different classes in the feature space. SVMs are effective in high-dimensional spaces and are known for their generalization ability and robustness against overfitting.

KNN:
K-Nearest Neighbors (KNN) is a simple, non-parametric, and lazy learning algorithm used for classification and regression. For a given input, it finds the k training examples that are closest in the feature space and makes predictions based on their outputs. KNN is easy to implement and understand, but can be computationally intensive as the dataset grows.

Large Language Models

Autoregressive Text Generation:
Autoregressive text generation models produce sequences of text by generating one token at a time, conditioning on the previously generated tokens. This approach enables the model to capture the context and structure of the input sequence, enabling coherent and contextually relevant text generation. GPT-3 and GPT-4 are prominent examples, leveraging vast amounts of training data and parameters to generate human-like text across diverse topics and styles.

Tokenization:
Tokenization is the process of converting input text into smaller units called tokens, which can be words, subwords, or characters. This process is crucial for preparing textual data for language models, allowing them to understand and analyze the structure, semantics, and syntax of the text. Various tokenization strategies exist, such as whitespace, sentence-piece, and byte pair encoding, each with its benefits and trade-offs.

Context Window:
The context window in NLP models refers to the amount of textual information, typically measured in tokens or words, that the model can consider simultaneously for processing and understanding. The size of the context window is crucial as it determines the extent to which a model can capture relationships, dependencies, and nuances in the text. For instance, Transformers have large context windows, enabling them to comprehend extensive contextual information and thereby produce more coherent and contextually relevant outputs.

Embeddings:
Embeddings are dense vector representations of discrete objects, such as words or tokens, in a continuous vector space. In the realm of NLP, word embeddings serve as a foundational

technique to represent and process textual data. They capture semantic and syntactic relationships between words by positioning similar words closer in the vector space. Pre-trained embeddings like Word2Vec, GloVe, and ELMo, or trainable embeddings in neural networks, facilitate models in understanding text, enabling them to perform various tasks like classification, translation, and generation by computing on the embedded vectors.

Transformers:
Transformers are a class of neural network architecture that have become the backbone of most modern large-language models (LLMs). Introduced by Vaswani et al. in 2017, Transformers use self-attention mechanisms to weigh and prioritize different parts of the input sequence, enabling superior performance in capturing long-range dependencies and relationships in text. Transformers' parallel processing capabilities also make them highly efficient and scalable.

Modern LLM Recipe:
Modern large language models (LLMs) are developed using a recipe that typically includes pre-training on a vast corpus of diverse text data, followed by fine-tuning on more specific tasks, rewards modeling using human feedback, and finally reinforcement learning to update the original LLM using the reward model.

Base LLM:
A base LLM is a large language model that is pre-trained on a general and extensive corpus of text data, serving as a starting point for more specific applications. GPT and LLAMA are examples. This base model captures a wide array of knowledge and linguistic structures, which can be fine-tuned or adapted to suit specialized tasks or domains, optimizing its performance for specific NLP challenges.

Supervised Fine-tuning:
Supervised fine-tuning is the process of refining a pre-trained language model on a labeled dataset for a specific task. This enables the model to adapt its generalized knowledge to the nuances and requirements of the target task, improving accuracy and effectiveness. It's a crucial step in tailoring base LLMs for applications like text classification, named entity recognition, and sentiment analysis. The most prevalent example is fine-tuning to respond to human inputs (a "chat" model).

Reward Modeling:
Reward modeling is a technique in reinforcement learning where a model is trained to predict the rewards of different actions based on human feedback. This approach enables the development of a reward function that guides the learning agent towards desired behaviors, helping it navigate complex and high-dimensional state and action spaces in various applications. For LLMs, humans rank the quality of outputs for a given prompt and

supervised learning is used to produce a model that predicts a "reward" which is higher for higher-quality outputs.

Reinforcement Learning (LLM):
Reinforcement learning (RL) is a machine learning paradigm where an agent learns by interacting with an environment, receiving rewards or penalties, and adjusting its actions to maximize cumulative rewards. For LLMs, the reward model (see above) assigns rewards to the outputs of the LLM and (some of) the LLM's model weights are adjusted to improve the outputs (typically using Proximal Policy Optimization or PPO).

In-context Learning & Prompt Engineering:
In-context learning refers to the ability of large language models to adapt to new information or tasks by conditioning on the provided context or prompts. Prompt engineering is the art of crafting input prompts that guide the model to generate desired outputs, effectively leveraging the model's knowledge and capabilities for specific tasks.

Retrieval-augmented Generation:
Retrieval-augmented generation (RAG) combines the strengths of retrieval-based and generative models. It retrieves relevant documents or text snippets from a corpus and uses them as additional context for generating responses. This enables the model to access external knowledge and produce more informed and accurate outputs, especially in open-domain question-answering tasks.

Vector Databases:
Vector databases store and manage high-dimensional vectors, typically embeddings produced by machine learning models. These databases enable efficient similarity search and retrieval of vectors, supporting applications like recommendation systems, information retrieval, and nearest neighbor search in high-dimensional space.

Chaining & Agents:
Chaining refers to connecting multiple models or agents in a sequence to solve complex tasks, with each model focusing on a sub-task or component of the problem. This modular approach enables the development of more versatile and capable systems, leveraging the specialized strengths of individual models or agents.

Evaluation of LLMs:
Evaluating large language models involves assessing their performance on specific tasks, considering aspects like accuracy, coherence, diversity, and bias. Evaluation can be quantitative, using metrics like BLEU, ROUGE, or F1 score, or qualitative, involving human judgment. Robust evaluation is crucial for understanding a model's strengths, weaknesses, and potential areas for improvement.

Programming and Tools

Python - A common language for AI/ML.

TensorFlow - A machine learning framework developed by Google.

PyTorch - An open-source machine learning library for Python, developed by Facebook's AI Research lab.

Scikit-Learn - A Python library for machine learning that supports supervised and unsupervised learning algorithms.

Pandas - A Python library providing data structures for efficiently storing large amounts of data.

Langchain - a framework for developing applications powered by language models. It enables applications that are context-aware, and able to reason and take action.

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.

Algorithm Audit & Assurance

Audit, Assurance, and Assessment:
Audit refers to the systematic examination of records, statements, or processes to ensure accuracy, compliance, and reliability. Assurance is the provision of an independent opinion on the quality and integrity of such information or processes, enhancing stakeholder confidence. Assessment is a broader evaluation of system performance, effectiveness, and risk, often encompassing audit and assurance activities.

Goals of an Audit/Assurance Engagement:
The primary goals of an audit/assurance engagement are to provide an independent and objective evaluation of statements, processes, or systems, ensuring their accuracy, compliance with applicable standards and regulations, and reliability. The engagement aims to enhance stakeholder trust and confidence by identifying and mitigating risks, discrepancies, and instances of non-compliance.

Independence Requirements:
Independence is a fundamental principle in auditing and assurance, requiring practitioners to remain unbiased and avoid conflicts of interest. More information on independence requirements can be found in professional standards, guidelines, and ethical codes issued by regulatory bodies and professional organizations, such as the International Ethics Standards Board for Accountants (IESBA).

Subject Matter and Relevant Criteria:
Subject matter refers to the information, system, process, or condition being audited or assured. Relevant criteria are the standards, frameworks, or benchmarks used to evaluate the subject matter. Selecting appropriate criteria is crucial for a meaningful and reliable audit or assurance conclusion.

Reasonable and Limited Assurance:
Reasonable assurance is a high but not absolute level of assurance provided by auditors regarding the absence of material misstatements or non-compliance. Limited assurance is a lower level of assurance, typically involving less extensive procedures and resulting in a less conclusive opinion.

Attestation and Direct Engagements:
In attestation engagements, a practitioner reports on a subject matter or an assertion about the subject matter that is the responsibility of another party. In direct engagements, the practitioner measures or evaluates the subject matter against criteria and presents the results without having to rely on an assertion from another party.

Suitable Criteria:
Suitable criteria are essential for a meaningful audit or assurance engagement. The criteria should be relevant to the subject matter, complete and reliable for consistent evaluation, neutral to avoid bias, and understandable to both the practitioner and the intended users of the report.

Audit Planning:
Audit planning involves defining the scope, objectives, and approach of the audit, identifying risks, and determining the necessary procedures and resources. Effective planning is

essential for conducting an efficient and thorough audit, ensuring that all relevant aspects are examined, and the audit objectives are achieved.

Material Misstatement:
A material misstatement is an error, omission, or misrepresentation in financial statements or other information that could affect the decisions of users or stakeholders. Identifying and addressing material misstatements are central to the integrity of an audit.

Audit Procedures:
Audit procedures are specific tests and inquiries conducted by auditors to gather evidence, evaluate risks, and form an opinion on the subject matter. They include inspection, observation, confirmation, recalculation, reperformance, analytical procedures, and inquiry.

Audit Evidence:
Audit evidence is the information collected during an audit to support the auditor's opinion. It includes records, documents, statements, and observations that are relevant and reliable for assessing compliance with the applicable criteria.

Audit Opinion:
An audit opinion is the conclusion expressed by the auditor based on the evaluation of audit evidence, regarding the accuracy, compliance, and reliability of the subject matter. Common types of audit opinions include unqualified (clean), qualified, adverse, and disclaimer of opinion.

ISAE 3000 (Revised):
ISAE 3000 (Revised) is an international standard for assurance engagements other than audits or reviews of historical financial information. It provides a framework for the planning, conducting, and reporting of assurance engagements, ensuring consistency, quality, and reliability across different types of assurance services.

Sarbanes-Oxley Act of 2002:
The Sarbanes-Oxley Act of 2002 is a US federal law aimed at enhancing corporate governance and accountability, particularly in financial reporting. It introduced stringent regulations for public companies, auditors, and corporate officers, including requirements for internal controls, auditor independence, and disclosure of financial information.

IAASB - ISQM:
The International Auditing and Assurance Standards Board (IAASB) develops international standards for auditing, assurance, and related services. The International Standard on Quality Management (ISQM) issued by IAASB provides guidelines for audit firms to manage and maintain quality in audit and assurance engagements.

Handbook of the International Code of Ethics for Professional Accountants:
This handbook, issued by the International Ethics Standards Board for Accountants (IESBA), contains the International Code of Ethics for Professional Accountants, providing principles and standards for ethical behavior, including integrity, objectivity, confidentiality, and professional behavior, applicable to all professional accountants.