The need for trust in AI has been of importance and one way of achieving it is through mitigating discrimination and bias in machine learning models throughout the AI application lifecycle. **This course will give you an overview** **on the concept of fairness which helps in building trust in AI and how "AI Fairness 360" open source toolkit can help you implement debiasing techniques to measure, understand and mitigate AI bias .** Learners will be provided an overview of AI fairness and bias concepts, how to measure bias in models and how to apply fairness algorithms to reduce bias. It will also walk you through a demo of working of "AI Fairness 360" open source tool kitand using this tool kit on a real-world use-case.

**By the end of this course, you will be able to:**

* Recognize the value of fairness in fostering trust in the world of AI
* Identify where unwanted bias comes from, various fairness metrics and unwanted bias mitigation methods to increase fairness and foster trust.
* Explain how AI Fairness 360 can help mitigate bias and discrimination in AI models
* Create fairer AI models by implementing bias mitigation techniques using AI Fairness 360 toolkit for a real world use case.

**This course is intended for:**

* Analytics Leaders
* Data Science Leaders and
* Practicing Data Scientists
* Machine Learning Engineers and
* AI specialists
* Anyone with an interest in AI Trust and bias mitigation concepts having the prerequisite knowledge that is mentioned below

**In order to be successful, you should have some knowledge of :**

* Data Science
* Machine Learning
* Python

**Estimated Learning Time:** ~ 4hr

* Total of 3 learning modules with 1.5 hours of self-paced video lectures
* 1 hour of additional reading
* 1 hour of Hands-on lab
* 0.5 hours of Quiz

In this module, you will:

* Recognize the need of Trustworthy AI
* Describe and differentiate various factors that can build trust in AI
* Appraise situations that require a focus on fairness
* Analyze where unwanted bias comes from
* Recognize methods to mitigate unwanted bias

To be successful in this module:

* Some foundational knowledge in Data science/ ML workflow and
* Evaluation metrics such as bias, accuracy etc is helpful

Definitions

**Bias**

A systematic error. In the context of fairness, we are concerned with ***unwanted bias*** that places privileged groups at systematic advantage and unprivileged groups at systematic disadvantage.

**Bias mitigation algorithm**

A procedure for reducing unwanted bias in training data or models.

**Classifier**

A model that predicts categorical labels from features.

**Explainer**

Functionality for providing details on or causes for fairness metric results.

**Fairness metric**

A quantification of unwanted bias in training data or models.

**Favorable label**

A label whose value corresponds to an outcome that provides an advantage to the recipient. The opposite is an unfavorable lable.

**Feature**

An attribute containing information for predicting the label.

**Group fairness**

The goal of groups defined by protected attributes receiving similar treatments or outcomes.

**In-processing algorithm**

A bias mitigation algorithm that is applied to a model during its training.

**Individual fairness**

The goal of similar individuals receiving similar treatments or outcomes.

**Instance weight**

A numerical value that multiplies the contribution of a data point in a model.

**Label**

A value corresponding to an outcome.

**Machine learning**

A general approach for determining models from data.

**Model**

A function that takes features as input and predicts labels as output.

**Post-processing algorithm**

A bias mitigation algorithm that is applied to predicted labels.

**Pre-processing algorithm**

A bias mitigation algorithm that is applied to training data.

**Privileged protected attribute**

A value of a protected attribute indicating a group that has historically been at systematic advantage.

**Protected attribute**

An attribute that partitions a population into groups whose outcomes should have parity. Examples include race, gender, caste, and religion. Protected attributes are not universal, but are application specific.

Provided below is a summary of the entire module. For your convenience, below you will see a heading that matches the content topics that were covered in this module, followed by a bulleted list of key concepts covered.

### Lesson 1: Need for Trustworthy AI

1. Recent “AI incidents” have been significant to prove that intelligent systems are prone to unforeseen and failures when they were deployed to the real world. Therefore it is essential to build trust in AI.

2. IBM Researchers believe that Trust in AI can be fostered through these pillars:

* **Explainability** is the ability of the AI model to explain how and why it arrived at a particular decision
* **Fairness** is the ability of the AI model to be free of bias in its decisions and to avoid unfair treatment of certain groups
* **Robustness** is the ability of the AI model to be safe and secure and not be vulnerable to any tampering or compromising the data they are trained on.
* **Transparency** is the ability to disclose information to increase the understanding of how an AI model or service was created and deployed
* **Governance** is the ability to direct, manage and control the AI activities throughout the AI lifecycle.

3. Procedural justice is perceived fairness of the rules and processes used to determine outcomes and Distributive justice is perceived fairness of the outcomes or resource allocations themselves.

### 

### **Lesson 2: AI Fairness**

4. Decision making supported by AI can have unwanted bias. Fairness in the context of machine learning is the ability of the AI model to mitigate unwanted bias in its decisions

**5. Unwanted bias places privileged groups at systematic advantage and unprivileged groups at systematic disadvantage**

**6. Unwanted bias comes from :**

* Problem misspecification
* Data engineering
* Prejudice in historical data
* Under - sampling

**7. Some Fairness terms:**

* Protected Attribute – an attribute that partitions a population into groups whose outcomes should have parity (ex. race, gender, caste, and religion)
* Privileged Protected Attribute – a protected attribute value indicating a group that has historically been at systemic advantage
* Group Fairness – Groups defined by protected attributes receiving similar treatments or outcomes
* Individual Fairness –Similar individuals receiving similar treatments or outcome
* Fairness Metric – a measure of unwanted bias in training data or models (ex: statistical parity difference vs equal opportunity difference)

**8. Bias Mitigation Methods : Common goal of introducing statistical independence between the predicted label and the protected attributes. This can be done at three places in the machine learning pipeline:**

* Pre - processing
* In - processing and
* Post - processing