An increasing concern is that rapid surge in the complexity and sophistication of AI models has evolved to such an extent that humans do not understand the complex mechanisms by which AI models make certain decisions. This understanding, referred to as “explainability” or “interpretability,” allows users to gain insight into the machine’s decision-making process. The need for trust in AI has been of importance and one way of achieving it is through creating explainable workflows throughout the AI application lifecycle. **This course will give you an overview on the concept of explainability which helps in building trust in AI and how "AI Explainability 360" open source toolkit can help you create explainable machine learning models.**

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

* Recognize the value of explainability in fostering trust in the world of AI
* Identify the applications of various explainability algorithms including the application of semi-automated data science and how they augment the data science pipeline to increase explainability and foster trust
* Explain how AI Explainability 360 can help you create explainable workflows
* Create explainable AI workflow using AI Explainability 360 toolkit for a real world use case

**This course is intended for:**

* Analytics Leaders
* Data Science Leaders
* Practicing Data Scientists
* Machine Learning Engineers
* 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: ~ 3 hr**

* Total of 3 learning modules with 1.5 hours of self-paced video lectures
* 30 min of additional reading
* 30 min of Bonus Exercise
* 30 min of Quiz Learning Outcomes and Prerequisites
* **In this module, you will:**
* Recognize the need for Trustworthy AI
* Describe and differentiate various factors that can build trust in AI
* Appraise situations that require a focus on AI explainability
* Recognize different methods of achieving explainability
* **To be successful in this module, you should have:**
* Prior knowledge is recommended in Data Science and Machine Learning
* Knowledge of evaluation metrics such as bias, accuracy, precision, and recall

## Michael Hind of IBM Research on AI Explainability (Optional)

#### Index of Videos:

1. Where did you go to school, and what is your main focus area?  
 2. How would you define explainability?  
 3. How does integrating explainability into AI help the end user?  
 4. Are there any use cases where explainability is being used in the marketplace right now?  
 5. Are there any major hurdles we need to get past to make explainability a reality?  
 6. Should black box models be totally avoided?  
 7. Do you think we understand what an explaination should entail?  
 8. Using ProtoDash as an example - when would prototyping data be useful?  
 9. Do you think explainable AI will change the way decisions are shared?  
 10. Where do you see explainability in the next 10 years?  
 11. How do you think soft skills play into explainable AI?  
 12. Why should people be excited about AI explainability?  
 13. Why would a software developer be interested in explanable AI?

Scroll down to watch each video

#### 1. Where did you go to school, and what is your main focus area? 2. How would you define explainability? 3. How does integrating explainability into AI help the end user? 4. Are there any use cases where explainability is being used in the marketplace right now? 5. Are there any major hurdles we need to get past to make explainability a reality? 6. Should black box models be totally avoided? 7. Do you think we understand what an explaination should entail? 8. Using ProtoDash as an example - when would prototyping data be useful? 9. Do you think explainable AI will change the way decisions are shared? 10. Where do you see explainability in the next 10 years? 11. How do you think soft skills play into explainable AI? 12. Why should people be excited about AI explainability? 13. Why would a software developer be interested in explanable AI?

## Module 1 Summary

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

**Lesson 1: The Need for Trusted AI**

1. Recent “AI incidents” have been significant to prove that intelligent systems are prone to unforeseen incidents 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

**Lesson 2: Explainable AI**

3. Explainable artificial intelligence (XAI) is a set of processes and methods that allows human users to comprehend and trust the output created by machine learning algorithms

**4. "Interpretability is the main motivation of Explainability." Interpretability and Explainability differ from each other in the following ways.**

* Interpretability means that the cause and effect can be determined whereas
* Explainability is the ability to help humans comprehend how machine learning models makes decisions

**5. Other motivations of Explainable AI are as follows**