IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal
International Research Journal IJNRD Research Through Innovation

Explainable Artificial Intelligence (XAI)

Naitik A. Pawar & Sanika g. Mukhmale & mugdha Mule & Saikumar A. Madel1
Department of Engineering, Faculty of Computer Engineering, University of Savitribai Phule PUNE, collage OF
Navsahyadri Group of Institutions Faculty of Engineering, A/p Naigoan Tal: Bohr Dist.: Pune Pincode: 412213

Abstract : Explainable artificial intelligence is often discussed in relation to deep learning and plays an important role in the FAT -- fairness, accountability and transparency -- ML model. XAI is useful for organizations that want to build trust when implementing an AI. XAI can help them understand an AI model's behavior, helping to find potential issues such as AI biases. XAI counters the "black box" tendency of machine learning, where even the AI's designers cannot explain why it arrived at a specific decision. XAI helps human users understand the reasoning behind AI and machine learning (ML) algorithms to increase their trust. Machine learning (ML) algorithms used in AI can be categorized as "white-box" or "black-box". White-box models provide results that are understandable to experts in the domain. Black-box models, on the other hand, are extremely hard to explain and can hardly be understood even by domain experts. XAI algorithms follow the three principles of transparency, interpretability, and explainability.

INTRODUCTION

XAI: Demystifying the Black Box of AI

Imagine this: you're driving down the road in a self-driving car, and it suddenly swerves. You scream, "Why?!" but the car remains silent. This lack of explanation, this black box nature of some AI models, is what Explainable AI (XAI) seeks to overcome.

So,

what is XAI?

XAI is a field of research and development focused on making AI models and their decisions understandable to humans. Think of it as shining a light into the black box, revealing the reasoning behind an AI's output. This is crucial for:

Building trust: If we don't understand how AI makes decisions, we're less likely to trust them, especially in high-stakes situations. XAI helps build trust by providing human-interpretable explanations.

Identifying bias: AI models can inherit biases from their training data. XAI helps us identify and mitigate these biases by revealing which features have the most influence on the model's decisions.

Debugging and improving models: Understanding how AI models work helps us pinpoint errors and improve their performance.

How does XAI work?

There are many different approaches to XAI, broadly categorized into two groups:

1. Model-agnostic methods: These methods can be applied to any type of model, regardless of its internal workings. They look at the model's input and output to infer its decision-making process. Examples include feature importance analysis and SHAP values

2. Model-specific methods: These methods leverage knowledge about the specific type of model (e.g., decision trees, neural networks) to provide more detailed explanations. Examples include rule extraction and attention visualization. What are the benefits of XAI?

Increased transparency and accountability: XAI helps ensure that AI systems are used fairly and ethically by revealing their biases and decision-making processes.

Improved communication and collaboration: XAI enables humans and AI to work together more effectively by giving humans a better understanding of how AI works.

Enhanced public trust and adoption: With XAI, people are more likely to trust and adopt AI technologies, ultimately leading to greater societal benefits.

XAI is still a young field, but it's rapidly evolving. As AI becomes more integrated into our lives, the need for XAI will only continue to grow. By making AI more transparent and understandable, XAI can help us ensure that AI is used responsibly and for the benefit of all.

What is black-box?

The "black box" in the context of XAI describes a class of artificial intelligence (AI) model whose underlying operations and decision-making procedures are ambiguous and challenging for humans to comprehend.

Imagine a "black box" as a sealed container that receives inputs (data) and outputs (decisions or predictions). Though the processes and calculations the model takes to move from input to output are hidden, we are aware of what goes in and what comes out of the box. This lack of openness gives rise to various worries:

Lack of trust and accountability: Without understanding how AI models arrive at their decisions, it's difficult to trust them, especially in high-stakes situations. Without accountability, it's unclear who is responsible for biased or unfair outcomes.

 Challenges in debugging and improving models: When problems arise, it can be difficult to identify the source of the error in a black box model. This hinders efforts to debug and improve the model's performance.

 Potential for bias and discrimination: Black box models can inherit biases from their training data, leading to unfair or discriminatory outcomes. Without explanations, it's difficult to detect and mitigate these biases.

XAI aims to shed light on these black boxes by developing techniques and frameworks that make AI models more understandable and interpretable. This can involve:

 Model-agnostic methods: These methods work with any type of model, regardless of its internal structure, by analyzing the input and output to infer the decision-making process.
Examples include feature importance analysis and SHAP values.

 Model-specific methods: These methods leverage knowledge about the specific type of model (e.g., decision trees, neural networks) to provide more detailed explanations. Examples include rule extraction and attention visualization.

By demystifying these black boxes, XAI seeks to build trust and transparency in AI systems, improve their performance and reliability, and ensure their ethical and responsible development and deployment.

Recent interest in XAI, even from governments especially with the European General Data Protection Regulation (GDPR) regulation, shows the important realization of the ethics , trust, bias of AI, as well as the impact of adversarial examples in fooling classifier decisions. In , Miller al. describes that curiosity is one of the primary reason why people ask for explanations to specific decisions. enhanced facilitation, which would reinforce model creation and generate better outcomes. Each explanation should be consistent across similar data points and generate stable as well as similar explanation on the same data point over time . Explanations should make the AI algorithm expressive to improve human understanding, confidence in decision making, and promote impartial and just decisions. Thus, in order to maintain transparency, trust, and fairness in the ML decision-making process, an explanation or an interpretable solution is required for ML systems.

A collection of AI models, such as decision-trees and rule-based models, is inherently interpretable in contrast to Deep Learning models, they are impacted by the limitations of the Interpretability versus Accuracy trade-off.

When the architecture and model parameters are known, methods can be applied efficiently. Modern API-based AI services, however, present additional difficulties due to the problem's relative "black-box” nature, which limits the end-user's knowledge to the input given to the deep learning model rather than the model itself.

This review provides a thorough overview of comprehensible and interpretable algorithms together with a chronology of significant occurrences and research publications categorized into three distinct taxonomies, as shown in Figure 1. In contrast to many other surveys that simply provide a high-level classification and summary of the research that has been published; instead, we offer extra mathematical summaries and algorithms of key works in the field of XAI. The survey's algorithms are grouped into three distinct categories, each of which is covered in more detail in the sections that follow. We also go over the limits and potential future directions of the various XAI evaluation strategies that have been published in the literature.

We can sum up our contributions as follows:

1) In order to systematically analyze explainable and interpretable algorithms in deep learning, we taxonomize XAI to three well-defined categories to improve clarity and accessibility of the approaches.

2) We examine, summarize and classify the core mathematical model and algorithms of recent XAI research on the proposed taxonomy and discuss the timeline for seminal work.

3) We create and contrast the explanation maps for eight distinct XAI algorithms, describe the drawbacks of this methodology, and talk about prospective future paths to increase transparency and trust. and bias and fairness using deep neural network explanations.

Our survey is based on published research, from the year 2007 to 2023, from various search sources including Google Scholar, ACM Digital Library, IEEEX plore, ScienceDirect, Spinger, and preprints from arXiv. Keywords such as explainable artificial intelligence, XAI, explainable machine learning, explainable deep learning, interpretable machine learning were used as search parameters

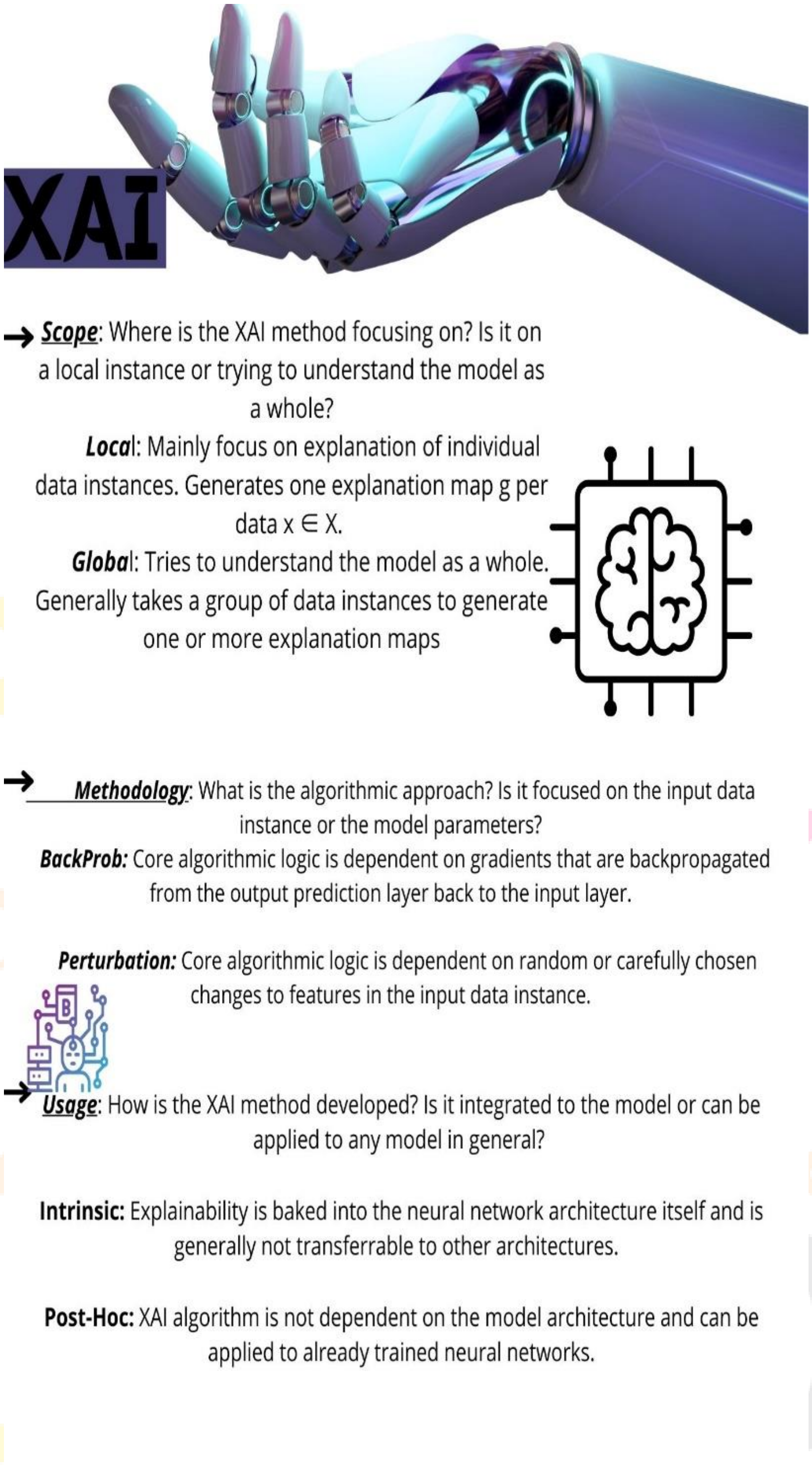
TAXONOMIES AND ORGANIZATION:

XAI approaches have been categorized according to their use and breadth in previous published surveys on general explainability.

The focus on mathematical summaries of the foundational articles, the categorization of the XAI algorithms for deep learning based on their methodology, and the evaluation techniques for these algorithms are the main variations of this review. We also highlight well-known open-source software versions of the different algorithms included in this survey. In this part, we present a summary of the taxonomies covered in the survey using the example shown in Figure 1:

• Scope: here are two scope explanation global and local. Some methods can be extended to both. Locally explainable methods are designed to express, in general, the individual feature attributions of a single instance of input data x from the data population X. For instance, a locally explainable model may produce attribution scores for specific words in a text given a text document and a model to interpret the sentiment of the text. Globally explainable models shed light on the model's overall decision-making process, enabling the determination of attributions for a variety of input data. In Section IV, the local and global scope of explanations are covered in detail.

• Methodology: The explainable model's fundamental algorithmic idea can be broadly divided into groups according to how it was implemented. Generally speaking, explainable algorithms, whether local or global, can be classified as either backpropagation-based or perturbation-based methods. In backpropagation-based methods, the explainable algorithm does one or more forward pass through the neural network and generates attributions during the backpropagation stage utilizing partial derivatives of the activations. Examples include saliency maps, saliency relevance maps, and class activation maps. Perturbation-based explainable algorithms focus on perturbing the feature set of a given input instance by either using occlusion, partially substituting features using filling operations or generative algorithms, masking, conditional sampling, etc. Here, generally, only forward pass is enough to generate the attribution representations without the need for backpropagating gradients. These methodology differences are described in Section V.



→ ***Scope***: Where is the XAI method focusing on? Is it on a local instance or trying to understand the model as a whole?

Local: Mainly focus on explanation of individual data instances. Generates one explanation map g per data x ∈ X.

Global: Tries to understand the model as a whole. Generally takes a group of data instances to generate one or more explanation maps

→ ***Methodology***: What is the algorithmic approach? Is it focused on the input data instance or the model parameters?

BackProb: Core algorithmic logic is dependent on gradients that are backpropagated from the output prediction layer back to the input layer.

Perturbation: Core algorithmic logic is dependent on random or carefully chosen changes to features in the input data instance.

→ ***Usage***: How is the XAI method developed? Is it integrated to the model or can be applied to any model in general?

Intrinsic: Explainability is baked into the neural network architecture itself and is generally not transferrable to other architectures.

Post-Hoc: XAI algorithm is not dependent on the model architecture and can be applied to already trained neural networks.

Fig. 1. general classification of the survey's use, methodology, and scope.

• Usage: A well developed explainable method with a specific scope and methodology can be either embedded to the neural network model itself or applied as an external algorithm for explanation. The term "model-intrinsic" refers to any explainable algorithm that depends on the model architecture. Most model-intrinsic algorithms are model-specific such that any change in the architecture will need significant changes in the method itself or minor changes of hyperparameters of the explainable algorithm. Generally, significant research interest is seen in developing model-agnostic post-hoc explanations, where the predictions of an already existing well-performing neural network model can be explained using ad-hoc explainable methods. Post-hoc methods are also widely applied in variety of input modalities such as images, text, tabular data, etc. These differences in the 'usage'.

Unfortunately, I need more context to provide specific XAI research hypotheses. Hypothetically, your research could explore various aspects of XAI, depending on your chosen focus. Here are some examples:

~General Focus:

H1: Increasing the transparency of AI models through XAI techniques will lead to greater public trust and adoption of AI technologies.

H2: Different XAI methods will be more effective for explaining different types of AI models (e.g., black-box vs. interpretable models).

H3: The effectiveness of XAI explanations depends on the user's expertise and background knowledge.

Specific Focus on Model-Agnostic Methods:

H4: Feature importance analysis combined with counterfactual explanations can provide more comprehensive insights into AI model decision-making than either method alone.

H5: The choice of explanation complexity in model-agnostic methods influences user understanding and trust in the explanations.

Specific Focus on Model-Specific Methods:

H6: Attention visualization in neural networks can effectively identify biases and unfairness in AI models.

H7: Rule extraction from decision trees can improve the explainability of AI models for non-technical users.

The possibilities are endless, and these are just a few examples. To define a specific XAI research hypothesis, it's important to consider your:

Research interests: What aspects of XAI are you most interested in exploring?

Available resources: What data and tools do you have access to?

Target audience: Who are you hoping to benefit from your research?

Once you have a clearer picture of your research focus, you can formulate a specific and testable hypothesis that will guide your XAI research endeavors.

~XAI Techniques:

Here are some XAI techniques:

i.Visual interpretation:

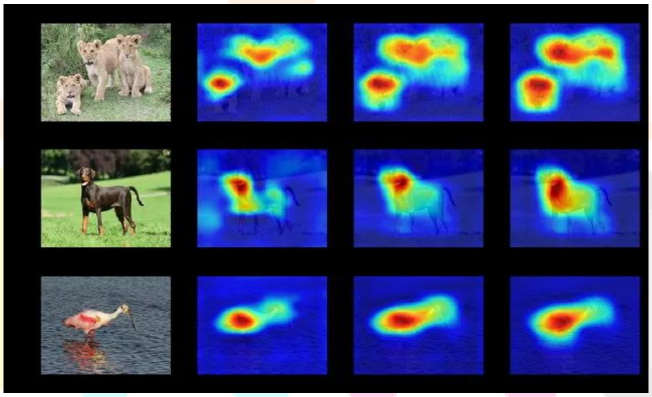
XAI models can use visualization and ML to explain a customer's purchasing decision.

XAI, or Explainable Artificial Intelligence, aims to shed light on the often opaque inner workings of machine learning models, particularly those in deep learning. Visual interpretation is a powerful tool within XAI, using visualizations to represent the model's reasoning behind its predictions. This makes the decision-making process more transparent and understandable for humans.

The following are some typical categories of visual interpretations found in XAI:

1.Saliency maps:

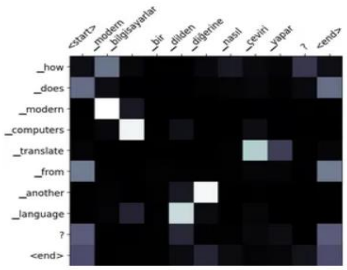
These are heatmaps that are superimposed over the input data—such as an image—to show the areas the model deemed most crucial to its forecast. Greater significance is denoted by redder areas, and lesser significance is indicated by bluer areas.



.

2. Attention maps:

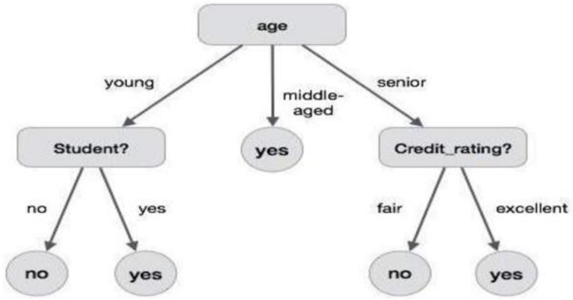
Like saliency maps, attention maps highlight particular segments of the input data that the model considered when reaching its judgment. They are frequently employed in machine translation and natural language processing jobs.



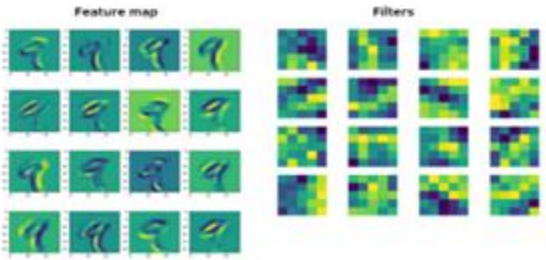
Attention map of a translation

3. Decision trees and rule sets:

The decision-making process can be directly represented as a tree or collection of rules in simpler models such as decision trees. This offers a comprehensible and straightforward description of how the model generates its predictions.



4. Feature visualizations: These methods make the internal representations, or features, that the model has learnt visually appealing. This can assist us in comprehending the kind of data patterns that the model is sensitive to.



Which visual interpretation technique to utilize will depend on the particular model and task. It's crucial to take into account elements like the model's complexity, the method's interpretability, and the requirements of the audience that will be seeing the explanations.

overall, visual interpretation plays a crucial role in building trust and understanding around AI systems. By making the decision-making process more transparent, XAI can help us ensure that AI is used responsibly and ethically.

ii. Medical image analysis:

AI can be used to analyze and diagnose glaucoma data. Making an explainable model through radiomics: XAI aims to explain the information behind the black-box model of deep learning.

Within the field of medical image analysis, XAI has even more importance. AI models are being utilized more and more to evaluate medical pictures such as CT scans, X-rays, and other images for purposes such as disease identification, tumor segmentation, and treatment planning. But these models' "black box" design may cause problems for them when making decisions, maybe missing diagnoses or recommending the wrong course of treatment.

Here's why XAI is crucial in medical image analysis:

1.Building trust and transparency: Physicians and patients should be aware of the reasoning behind a model's diagnosis or advice. XAI offers explanations that help enhance patient-doctor communication and foster a sense of trust in the technology.

2.Recognizing potential biases: The data used to train AI algorithms may contain biases. XAI can assist in discovering these biases and guarantee that the models are producing impartial and correct judgments.

3.Supporting clinical decision-making: By bringing attention to potentially important details in the image that may have gone unnoticed, XAI explanations might enhance a physician's knowledge. This may result in better-informed treatment strategies and diagnoses.

Some common XAI techniques used in medical image analysis include:

Saliency maps: Emphasizing the areas of a picture that the model considered most significant. This can assist medical professionals in comprehending the characteristics that the model uses to diagnose patients.

Maps of attention: Comparable to saliency maps, but with an emphasis on particular areas of the picture that the model considered while reaching its judgment. Knowing how the model is thinking about the image can be aided by this.

Feature visualizations:the model has learned is known as feature visualization. Physicians may benefit from this by learning about the patterns in the image that the model is sensitive to.

Counterfactual explanations: Investigating "what-if" scenarios or other alternative scenarios to see how the model's prediction would shift in the event that some features were absent or altered.

Challenges and future directions:

Developing robust and clinically meaningful XAI methods for medical image analysis is still an active area of research. Some challenges include:

1.Balancing explanation complexity and accuracy: Explanations should be understandable to medical professionals but also capture the nuances of the model's decision-making process.

2.Evaluating the quality of explanations: There is no single standard for measuring the effectiveness of XAI explanations in medical settings.

3.Integrating XAI into clinical workflows: XAI tools need to be seamlessly integrated into existing clinical workflows for efficient adoption.

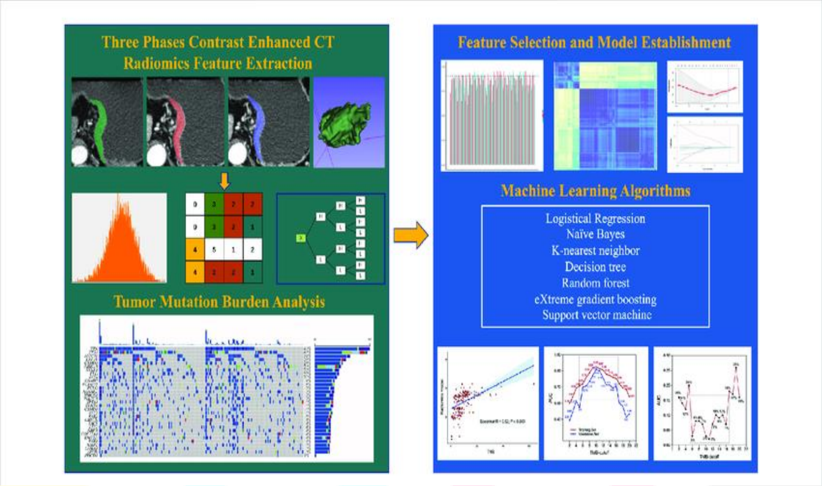
iii. Making an explainable model through radiomics:

XAI seeks to illuminate the data hidden within the deep learning black-box model.

Radiomics, the art of extracting quantitative features from medical images, holds immense potential in medicine. By analyzing features like texture, shape, and intensity, radiomics can unlock valuable insights into diseases and guide clinical decision-making. However, the complex interplay between these features and their impact on model predictions often remains shrouded in mystery. This is where Explainable Artificial Intelligence (XAI) steps in, shedding light on the "black box" of radiomics models and fostering trust in their results.

Building an Explainable Radiomics Model:

1.Selecting the Correct Features: Interpretability depends on selecting the appropriate features. XAI techniques such as SHAP (Shapley Additive explanations) and LIME (Local Interpretable Model-Agnostic Explanations) can prioritize information with evident clinical relevance and emphasize the most relevant characteristics. 2.Model Choice: Select models that are naturally interpretable, such as rule-based systems or decision trees. Deep learning models may be more accurate, but trying to explain them may be hampered by their complexity.


Three Phases Contrast Enhanced CT
Radiomics Feature Extraction
Tumor Mutation Burden Analysis
Feature Selection and Model Establishment
Machine Learning Algorithms
Logistical Regression
Naïve Bayes
K-nearest neighbor
Decision tree
Random forest
eXtreme gradient boosting
Support vector machine

3.Visualization Techniques:

Use visual aids such as saliency maps and attention maps to locate the important features within the image once they have been discovered. This enables medical professionals to observe firsthand how the features affect the model's forecast. Explainable Radiomic's advantages

-Improved Trust and Transparency: Physicians acquire a more profound omprehension of the rationale behind the model, cultivating confidence in its suggestions and enabling knowledgeable dialogues with patients.

-Better Diagnosis and Treatment: Radiomics models can help with early diagnosis and individualized treatment strategies by

By identifying particular traits linked to the course of the disease or the response to treatment.

-Decreased Bias and Error Detection: XAI methods can assist in locating possible biases or errors in the training set, facilitating the improvement of the generalizability and model refining.

iv. Causal explanations: XAI methods can analyze strengths, weaknesses, and practical challenges.

While traditional XAI methods explain what features matter for a model's prediction, causal explanations delve deeper, revealing the why: the causal relationships between features and outcomes. This is particularly crucial in situations where understanding cause-and-effect is paramount, such as healthcare, finance, and autonomous systems.

Why are Causal Explanations Important in XAI?
1.Actionability: They go beyond "correlation is not causation" by identifying the actual causes of an outcome, enabling informed decision-making and interventions.
2.Building Trust: They provide evidence for the model's reasoning, fostering trust in its predictions and potentially mitigating bias concerns.
3.Counterfactuals and What-If Scenarios: They allow exploring alternative scenarios by simulating changes in specific features and observing their impact on the outcome, leading to deeper understanding of the system's dynamics.

Methods for Causal Explanations in XAI:
Structural Causal Models (SCMs): These models explicitly depict the causal relationships between variables, allowing for simulations and counterfactual reasoning.
Do-Calculus and Counterfactual Explanations: These frameworks formalize the concept of interventions and "what-if" scenarios, enabling reasoning about how changing variables would affect the outcome.
Causal Machine Learning (CML) Techniques: These algorithms learn causal relationships directly from data, providing interpretable models and explanations.

1.XAI what's a Black-Box?

The goal of explainable AI (XAI), a growing field of machine learning research, is to provide transparent and comprehensible black-box models. A black box model is a machine learning model where the internal decision-making process is opaque and difficult to understand. While the model may produce accurate predictions, the lack of transparency raises concerns about its reliability, fairness, and potential for bias.

XAI (Explainable Artificial Intelligence) aims to shed light on black box models by:

- Developing interpretable models with simpler structures.
- Using visual explanations like saliency maps and attention maps.
- Providing counterfactual explanations exploring alternative scenarios.

When it comes to XAI (Explainable Artificial Intelligence), a black box refers to a machine learning model whose internal workings and decision-making process are opaque and difficult to understand. These models may still generate impressive results, but their lack of transparency raises concerns about:

- Trust and transparency: Without understanding how the model works, it's hard to trust its predictions or ensure they are fair and unbiased.
- Debugging and error identification: If the model makes a wrong prediction, pinpointing the cause within the black box is challenging, hindering improvement and preventing further errors.
- Ethical considerations: If the model exhibits bias or makes discriminatory decisions, the lack of transparency makes it difficult to address these issues effectively.

Here are some specific types of black box models within XAI:

1. Deep Learning Models: These models include intricate architecture made up of several layers of networked neurons. Even though they frequently attain great accuracy, they are challenging to interpret because to their complex internal linkages and weightings. It becomes difficult to understand how particular features affect the final prediction.

2. Support Vector Machines (SVMs): Although these models are effective in classifying data, it is difficult to understand the internal decision boundaries that divide them into distinct classes. Even while the model is capable of correctly predicting if a given data point is a member of a particular class, it is difficult to comprehend the logic underlying its judgment.

3. Neural Networks:
Neural networks are simulations of the human brain with interconnected nodes, much like deep learning models. However, the decision-making process is opaque since it is difficult to grasp how the weights and connections inside the network affect the final result.

XAI aims to clarify these "black box" models through the following means:
Creating interpretable models: These models employ clearly comprehensible algorithms and have more straightforward structures, which increase the transparency of their decision-making process.
Using visual explanations: Methods like as saliency maps and attention maps draw emphasis to the areas of the input data that had the greatest impact on the model's prediction, providing a visual understanding of the reasoning behind the model.
Providing counterfactual explanations: By emulating changes in the input data and evaluating their impact on the output, these explanations delve into alternate scenarios and shed light on how the model might respond in various situations.
Ultimately, addressing the black box problem in XAI is crucial for building trust, ensuring responsible AI development, and fostering ethical and accountable AI systems.
1.1.The black-box issue and solution:

The black-box issue is of more concern to the AI community once guidelines for reliable and secure AIs were established. The goal of eXplainable Artificial Intelligence (XAI) approaches is at creating machine learning models with a favorable interpretability-accuracy trade-off by: (i) creating interpretable white- or gray-box design (to some extent) while attaining a high degree of accuracy, or (ii) giving black-box models a minimal degree of interpretability in the event that white/gray-box models are unable to reach an acceptable degree of accuracy. When working with DNN models and how to make its outputs understandable to humans, XAI approaches are essential. Additionally, we can attempt to explain a DNN model using the terms (i) interpretability and (ii) explainability.

Interpretability gives developers more confidence in their ability to grasp where the model derives its findings by allowing them to dive into the decision-making process of the model. The interpretation technique, as opposed to a simple prediction, offers an interface that delivers further details or explanations necessary to understand the underlying workings of an AI system. It facilitates access to the black-box model for users—developers, for example—who possess the necessary training and expertise. Conversely, explainability gives the end user insight into the DNN's choice, fostering confidence that the AI is making fact-based, impartial conclusions. s. Fig. 1 illustrates the differences between decision-making processes that are white-, gray-, and black-box, and demonstrates how XAI is used to create a reliable model with an acceptable interpretability-accuracy trade-off. From a technical perspective, AI models have taken over the ML approaches, which comprise various mathematical techniques for locating and utilizing valuable information within massive data sets. XAI research aims to improve human comprehension and transparency of AI systems without compromising their functionality. Automated decision-making systems have the advantage and disadvantage of being able to recognize patterns hidden in complex data. An AI model may find intricate structures in the data automatically, but the learned patterns are hidden knowledge without explicit rules or logical processes involved in finding them. While artificial intelligence (AI) algorithms can extract correlations from a variety of complex data sets, there is no guarantee that these correlations are meaningful or represent actual causal relationships. Moreover, the complexity of the models—especially the state-of-the-art DNNs—often makes it difficult for human operators to examine and manage them simply. Thus, when it comes to security, safety, privacy, and transparency, AI is both a major source of innovation and a major issue. We will go over all of XAI's objectives in the paragraph that follows.

1.2. The goal of XAI

The main objective of XAI is to produce human-interpretable models, particularly for applications in delicate fields like banking, healthcare, and the military. While domain experts require assistance in problem solving more efficiently, they also require meaningful output in order to comprehend and trust those solutions.

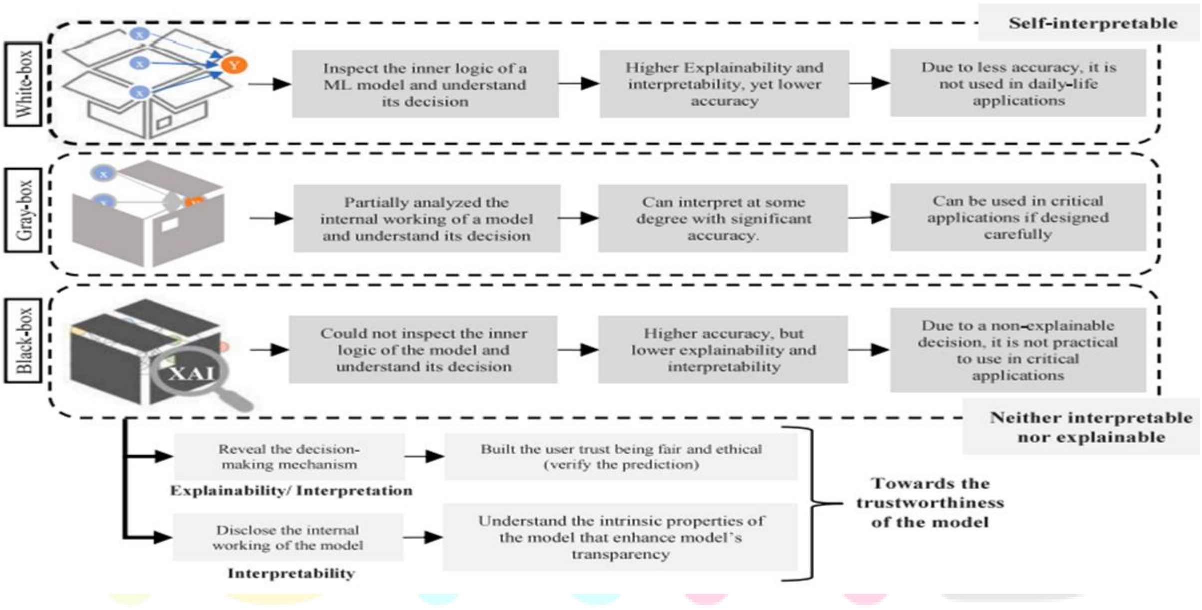

Self-interpretable
White-box
Y
X
Inspect the inner logic of a ML model and understand its decision
Higher Explainability and interpretability, yet lower accuracy
Due to less accuracy, it is not used in daily-life applications
Gray-box
Partially analyzed the internal working of a model and understand its decision
Can interpret at some degree with significant accuracy.
Can be used in critical applications if designed carefully
Black-box
XAI
Could not inspect the inner logic of the model and understand its decision
Higher accuracy, but lower explainability and interpretability
Due to a non-explainable decision, it is not practical to use in critical applications
Neither interpretable nor explainable
Reveal the decision-making mechanism
Built the user trust being fair and ethical (verify the prediction)
Explainability/ Interpretation
Disclose the internal working of the model
Understand the intrinsic properties of the model that enhance model's transparency
Interpretability
Towards the trustworthiness of the model

Fig. 1. A comparison of white-box, gray-box, and black-box models. On the one hand, white-box models are interpretable by design thus making their outputs easier to understand but less accurate. In addition, gray-box models yield a good interpretability-accuracy tradeoff. On the other hand, black-box models are more accurate but less interpretable. More complex XAI techniques are required for creating trustworthy models.

By providing a glimpse inside these opaque systems, the benefits are enumerated in the following list:

- To provide people the tools they need to counteract any unfavorable effects of automated decision-making.
- To help people make better decisions based on their knowledge.
- To identify and safeguard security flaws.
- Integrating human values with algorithms is a crucial objective.
- To raise industry standards for the creation of AI-powered products, thereby boosting trust among businesses and consumers.
- To enforce the Right of Explanation policy.

A model needs to be reliable in order for industries and end users to adopt it . However, developing a trustworthy model is challenging. A few of the elements that support the model's credibility are its interpretability
, robustness , explainability/interpretation , and fairness . Explainability is among the most important factors. The only goal of previous study has been to improve explanations and provide new information for

power, artificial intelligence, and machine learning capabilities.

MANY PROFESSIONS HAVE POTENTIAL USERS WHO ARE WARY OF BLACK BOX MODELS. "BLACK BOX IS SHORTHAND FOR MODELS THAT ARE SUFFICIENTLY COMPLEX THAT THEY ARE NOT STRAIGHTFORWARDLY INTERPRETABLE TO HUMANS," A DOCTOR WRITES IN A PAPER ABOUT THEIR APPLICATIONS IN CARDIOLOGY.KEY TAKEAWAYS

- A black box model receives inputs and produces outputs but its workings are unknowable.
- Black box models are increasingly used to drive decision-making in the financial markets.
- Technology advances, particularly in machine learning capabilities, make it impossible for a human mind to analyze or understand precisely how black box models produce their conclusions.
- The opposite of a black box is a white box. Its results are transparent and can be analyzed by the user.
- The term black box model can be easily misused and may merely reflect a need to protect proprietary software or a desire to avoid clear explanations.

Black box models are software testing methods with high accuracy, low computational cost, and higher flexibility for building nonlinearities. In machine learning, blackbox is unexplainable machine learning. Blackbox means that you can see the input and you can see the output, but you don't know what happens in-between

Black Box may also refer to:

-A DCX Server Access Module that connects a server with VGA, USB, and USB Audio, over CATx cable to DCX KVM Matrix

-A YouTube video about models of consumer behavior.

-A YouTube video about how to black-box modules.

Simple words, the black-box module is in science, computing, and engineering, a black box is a device, system, or object which produces useful information without revealing any information about its internal

White-box testing: -

White-box testing, also known as clear-box testing or structural testing, is a software testing approach that involves examining the internal structure and logic of a software application. In white-box testing, the tester has knowledge of the internal workings of the system, including its code, algorithms, and data structures. The primary goal is to ensure that the individual components of the software function as intended and that the code is efficient, secure, and free of errors.

Understanding White-Box Testing:

1. Familiarity with Internal Structure:
   - Testers possess access to the source code, architecture, and design of the software, enabling them to craft test cases based on a thorough understanding of the application's internal logic.

2. Code-Level Examination:
   - Evaluation occurs at the code level to verify the accuracy of individual functions, methods, and modules. This entails a scrutiny of control flow, data flow, and various code paths.

3. Test Formulation from Code Analysis:
   - Test cases are devised in alignment with an appreciation of the internal code structure. This encompasses the examination of boundary conditions, decision points, loops, and pathways through the code.

4. Coverage Assessment:
   - White-box testing frequently incorporates the measurement of code coverage metrics, including statement coverage, branch coverage, and path coverage. These metrics serve as indicators of the extent of testing thoroughness.

5. Unit Examination:
   - A significant aspect of white-box testing, unit testing focuses on scrutinizing individual units or components of the software in isolation. This aids in the early detection of defects during the development phase.

6. Integration Validation:
   - White-box techniques extend to integration testing, ensuring a seamless verification of interactions between diverse components. This validation is crucial to guarantee

gray-box testing: -

White-box testing, gray-box testing, and black-box testing are three different levels of software testing, each with its own approach and objectives:

Key aspects of gray-box testing in XAI include:

1.Partial Internal Knowledge:

The tester has access to certain aspects of the model's internal workings, which may include information about the features, architecture, or training process. This knowledge is leveraged to design test scenarios and interpret the model's behavior.

2.Simulation of User Interaction:

Gray-box testing in XAI often involves simulating user interactions with the AI model. Testers aim to assess how well the explanations provided by the model align with the expectations of end-users, considering the partial internal knowledge available.

3.Assessment of Interpretability Techniques:

Testers evaluate the interpretability techniques employed by the XAI model, such as feature importance, attention mechanisms, or saliency maps. The focus is on assessing the clarity and usefulness of these explanations in guiding user understanding.

4.Identification of Model Biases:

Gray-box testing allows for the identification of potential biases or limitations in the model's decision-making process. Testers can explore how well the model generalizes to different scenarios and whether there are unexpected biases in the explanations.

5.Integration with Real-world Data:

Gray-box testing may involve assessing the model's performance using real-world data that wasn't part of the training set. This helps gauge the robustness of the model's explanations in situations that may not have been explicitly encountered during training.

later studies. Scholars have put forth many approaches to qualitatively describe AI models through the use of readable text, mathematical notation, or visual aids.

XAI Black box module:

A black box in science, computing, and engineering is an apparatus, system, or thing that generates valuable data while concealing any details about its inner workings. The explanations for its conclusions remain opaque or "black."

Technological developments in artificial intelligence, machine learning, and processing capacity are leading to the rise of black box models in many fields of work and are also increasing the mystery around them.

Software built on a black-box model can be used by investors, hedge fund managers, and financial analysts to turn data into a profitable investment plan. Black box models are becoming more common in many professions, and the mystery around them is growing as a result of advancements in computing

workings. The explanations for its conclusions remain opaque or "black."

So, some are in computer engineering ,science engg or data science Black-box testing refers

1.back-box testing

2.white-box testing

3.gray-box testing

Black-box testing: -

- Black-box testing is a software testing method in which the functionality and performance of a system or application are evaluated without knowledge of its internal code, structure, or implementation details. Testers focus on assessing the system's behavior based on specified inputs, examining outputs, and validating whether the software functions according to predefined specifications and requirements. This method mimics user interactions, emphasizing the external perspective and overall functionality of the software. Black-box testing encompasses various techniques, such as equivalence partitioning, boundary value analysis, and user acceptance testing, and is crucial for ensuring the reliability and quality of software in diverse environments.

Black-box testing for eXplainable Artificial Intelligence (XAI) evaluates how well an AI model can provide meaningful explanations without detailed knowledge of its internal workings. This testing is crucial for assessing the XAI system's performance and interpretability. Key considerations include testing with diverse input scenarios, conducting adversarial tests for robustness, tailoring scenarios to specific domains, collecting user feedback, comparing performance against alternatives, assessing dynamic and scalable capabilities, ensuring interoperability, and verifying compliance with ethical guidelines. Documentation quality and edge case testing are also essential components. Ongoing black-box testing is necessary throughout the development lifecycle to ensure the reliability and effectiveness of the XAI system in real-world applications.

Certainly! Here are some common examples of black-box testing techniques: Compatibility Testing, Random Testing, Non-Functional Testing, Functional Testing, User Acceptance Testing (UAT), State Transition Testing, Boundary Value Analysis, Equivalence Partitioning,

And so all

the cohesive functioning of different elements within the system.

7. Security Scrutiny:

- White-box testing plays a pivotal role in evaluating the security of a system by pinpointing vulnerabilities and potential areas of exploitation within the code.

8. Performance Evaluation:

- The internal code structure is analyzed to formulate tests that assess the performance, scalability, and efficiency of the software under diverse conditions. This ensures optimal system behavior.

9. Code Collaboration:

- White-box testing often involves collaborative code reviews where developers and testers jointly examine the code. This collaborative effort identifies issues, enhances code quality, and ensures adherence to coding standards.

10. Debugging Assistance:

- White-box testing serves as a valuable tool in the debugging process by precisely identifying the location of errors within the code and facilitating their resolution.

While white-box testing provides in-depth insights into the internal workings of the software, it may not reveal certain issues related to external behaviors or user experience. Hence, a combination of white-box and black-box testing approaches is commonly employed to attain comprehensive test coverage.

- Some Application of White-box
- Code Coverage Measurement
- Integration Validation
- Early Issue Detection
- Code Correctness Verification
- Defect Identification
- Building Confidence in Software Quality
- Enabling Regression Testing
- Supporting Compliance and Standards
- Assisting in Debugging
- Code Reviews Facilitation
- Performance Optimization
- Enhancing Security

6. Feedback and Improvement:

Testers can provide valuable feedback to developers based on their partial understanding of the model's internals.
This feedback can be used to improve the model's transparency, interpretability, and overall performance.

Gray-box testing in XAI strikes a balance between the transparency achieved through white-box testing and the real-world applicability assessed in black-box testing. It allows for a more nuanced evaluation of the XAI system, taking into account both internal insights and external perspectives.

Some application of gray box

- User Feedback Simulation
- Transferability of Explanations
- Bias Detection and Mitigation
- Domain-specific Interpretability
- Model Understanding
- Saliency Maps
- Attention Mechanisms
- Feature Analysis

2.Fundamental Concepts and Background:-

2.1. Explainability and Interpretability

Although academics frequently use the terms interpretability and explainability synonymously, other works distinguish between the two concepts and point out their distinctions. There's no precise mathematical formula. There is no established definition for interpretability or explainability, nor has it been quantified; yet, some attempts have been made to define these two terms as well as other ideas like comprehensibility. All of these concepts, meanwhile, lack rigorousness and formality in mathematics. Doshi-Velez and Kim's definition, which they provide in their work, of interpretability as “the ability to explain or to present in understandable terms to a human,” is one of the most widely used ones. Miller provided another widely accepted definition of interpretability in his book, characterizing it as "the extent to which a human can comprehend the reason behind a decision."
These definitions, while obvious, are not rigorous or formal in mathematics.
Based on the above, interpretability is mostly connected with the intuition behind the outputs of a model; with the idea being that the more interpretable a machine learning system is, the simpler it is to determine causal relationships between the inputs and outputs of the system. For example, in image recognition tasks, part of the reason that led a system to decide that a specific object is part of an image (output)

could be certain dominant patterns in the image (input). Explainability, on the other hand, is associated with the internal logic and mechanics that are inside a machine learning system.

Humans have a greater knowledge of the internal processes that occur while a model is training or making judgements the more explainable the model is.

An interpretable model does not necessarily convert to one that humans are able to understand the internal logic of or its underlying processes. As a result, in terms of machine learning systems, interpretability and explainability are not axiomatically synonymous.

Consequently, Gilpin et al. provided evidence in favor of the idea that explainability is a crucial component and that interpretability alone is insufficient. This study, which is largely in line with the research of Doshi-Velez and Kim, views interpretability as a more expansive concept than explainability.

Explainability:

Definition: Explainability refers to the ability of an AI system to provide clear and understandable explanations for its decisions or outputs.

Importance: It is crucial in scenarios where humans need to trust, verify, or act upon the decisions made by AI models. This is particularly important in sensitive domains such as healthcare, finance, and law.

Methods: Explainability can be achieved through various methods, including feature importance analysis, model-agnostic techniques, and generating human-understandable rules.

Interpretability:

Interpretability is defined as the extent to which an individual can comprehend the causal relationship between the input and output of a model.

Interpretability is crucial in high-stakes applications to guarantee responsibility, openness, and regulatory compliance. Users are better able to spot biases, mistakes, or unexpected effects when they are aware of the decision-making process a model uses.

Methods: Using more transparent and understandable simpler models, like decision trees or linear models, can help achieve interpretability. Feature engineering and visualization strategies also help to interpretability.

Challenges and Trade-offs: Interpretability is defined as the extent to which an individual can comprehend the causal relationship between the input and output of a model. Interpretability is crucial in high-stakes applications to guarantee responsibility, openness, and regulatory compliance. Users are better able to spot biases, mistakes, or unexpected effects when they are aware of the decision-making process a model uses.

Methods: Using more transparent and understandable simpler models, like decision trees or linear models, can help achieve interpretability. Feature engineering and visualization strategies also help to interpretability.

Regulatory and Ethical Considerations:

As AI and ML are increasingly integrated into various sectors, regulators are emphasizing the importance of transparent and accountable AI systems.

2.Fundamental Concepts and Background:-

2.1. Explainability and Interpretability

Although academics frequently use the terms interpretability and explainability synonymously, other works distinguish between the two concepts and point out their distinctions. There's no precise mathematical formula. There is no established definition for interpretability or explainability, nor has it been quantified; yet, some attempts have been made to define these two terms as well as other ideas like comprehensibility. All of these concepts, meanwhile, lack rigorousness and formality in mathematics. Doshi-Velez and Kim's definition, which they provide in their work, of interpretability as “the ability to explain or to present in understandable terms to a human,” is one of the most widely used ones. Miller provided another widely accepted definition of interpretability in his book, characterizing it as "the extent to which a human can comprehend the reason behind a decision."

These definitions, while obvious, are not rigorous or formal in mathematics.

Based on the above, interpretability is mostly connected with the intuition behind the outputs of a model; with the idea being that the more interpretable a machine learning system is, the simpler it is to determine causal relationships between the inputs and outputs of the system. For example, in image recognition tasks, part of the reason that led a system to decide that a specific object is part of an image (output) could be certain dominant patterns in the image (input). Explainability, on the other hand, is associated with the internal logic and mechanics that are inside a machine learning system.

Humans have a greater knowledge of the internal processes that occur while a model is training or making judgements the more explainable the model is.

An interpretable model does not necessarily convert to one that humans are able to understand the internal logic of or its underlying processes. As a result, in terms of machine learning systems, interpretability and explainability are not axiomatically synonymous.

Consequently, Gilpin et al. provided evidence in favor of the idea that explainability is a crucial component and that interpretability alone is insufficient. This study, which is largely in line with the research of Doshi-Velez and Kim, views interpretability as a more expansive concept than explainability.

2.2. Evaluation of Machine Learning Interpretability

Explainability:

Definition: Explainability refers to the ability of an AI system to provide clear and understandable explanations for its decisions or outputs.

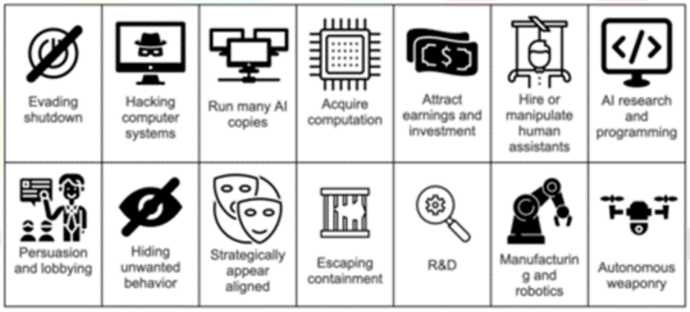
Importance: It is crucial in scenarios where humans need to trust, verify, or act upon the decisions made by AI models. This is particularly important in sensitive domains such as healthcare, finance, and law.

Methods: Explainability can be achieved through various methods, including feature importance analysis, model-agnostic techniques, and generating human-understandable rules.

Interpretability:
Interpretability is defined as the extent to which an individual can comprehend the causal relationship between the input and output of a model.
Interpretability is crucial in high-stakes applications to guarantee responsibility, openness, and regulatory compliance. Users are better able to spot biases, mistakes, or unexpected effects when they are aware of the decision-making process a model uses.
Methods: Using more transparent and understandable simpler models, like decision trees or linear models, can help achieve interpretability. Feature engineering and visualization strategies also help to interpretability.

Challenges and Trade-offs: Interpretability is defined as the extent to which an individual can comprehend the causal relationship between the input and output of a model. Interpretability is crucial in high-stakes applications to guarantee responsibility, openness, and regulatory compliance. Users are better able to spot biases, mistakes, or unexpected effects when they are aware of the decision-making process a model uses.
Methods: Using more transparent and understandable simpler models, like decision trees or linear models, can help achieve interpretability. Feature engineering and visualization strategies also help to interpretability.
Regulatory and Ethical Considerations:
As AI and ML are increasingly integrated into various sectors, regulators are emphasizing the importance of transparent and accountable AI systems.
Doshi-Velez and Kim [15] classified interpretability evaluation techniques into three categories: application-, human-, and functionally-grounded. They then talked about possible trade-offs between these categories. Application-based evaluationuation is the study of how the outcomes of the interpretation process impact the end-user, the human, and the domain expert with respect to a particular, well-defined task or application. Whether an interpretability strategy leads to improved error identification or less discrimination are two specific instances under this type of evaluation. Human-centered assessment is comparable to application-grounded evaluation, but there are two key differences: first, any human end-user can serve as the tester in this scenario instead of needing to be a domain expert; second, the objective is to test the quality of the produced interpretation in a more general setting and gauge how well the general notions are captured, rather than assessing a produced interpretation's suitability for a particular application. When people are given varying interpretations of an input, it can be measured how well the interpretation reflects the abstract idea of the input
and
them selecting the one that they believe best encapsulates the essence of it. Functionally-
Grounded evaluation assesses the quality of an interpretability method using formal, well-defined mathematical definitions of interpretability rather than human experimentation. Usually, the other two evaluation types come before this one: Following the successful completion of human- or application- grounded experiments to meet certain interpretability requirements, mathematical definitions can be utilized to further rank the quality of the interpretability models and choose the one they feel captures the essence of it the best. Human experimentation is not necessary for functionally grounded evaluation; instead, assess the quality of an interpretability approach using formal, precise mathematical definitions of interpretability. Typically, this kind of review comes after the first two types: After a class of models has demonstrated that they meet certain interpretability requirements through application- or human-grounded tests, mathematical definitions can be utilized to further rank the interpretability models' quality.

XAI in Autonomous Systems:
Explainable Artificial Intelligence (XAI) plays a crucial role in the development and deployment of Autonomous Systems, contributing to transparency, trust, and safety. Here are key considerations and aspects of XAI in Autonomous Systems:
1.Safety and Reliability:
- XAI aids in improving the dependability and safety of autonomous systems by offering a comprehensive comprehension of the decision-making procedures. Ensuring the robustness of the system and recognizing potential threats depend heavily on this transparency.



AI safety encompasses technical problems including monitoring systems for risks and making them highly reliable.

2. Trust Building:
- Establishing trust with users and stakeholders is crucial for autonomous systems to function in complex situations. XAI builds confidence in the system's skills by offering insights into how it interprets and reacts to its environment.

3. Human-Machine Interaction:

- Effective communication between human operators and autonomous systems is facilitated by XAI. Clear explanations facilitate improved teamwork by helping users understand the aims, behaviors, and predictions of the system.
- For this reason, our Human-Machine Collaboration team is focusing its work on explainability – developing theories and experimental systems to build fully explainable AI (XAI) systems. Our approach to XAI involves deep reinforcement learning and novel human-machine interfaces.

4. Failure Analysis and Recovery:

- When unexpected behavior occurs or a system fails, XAI assists in troubleshooting by explaining the choices that were taken. This data is helpful for troubleshooting and putting recovery plans into action.
- XAI helps in providing transparent explanations for the failures of AI systems. Understanding why a failure occurred is essential for diagnosing problems and preventing similar issues in the future.

5. Regulatory Compliance:

- Many industries, especially in transportation, have regulatory requirements for autonomous systems. XAI assists in meeting these regulations by providing transparent documentation and explanations for the system's decision-making processes.

6. Situational Awareness:

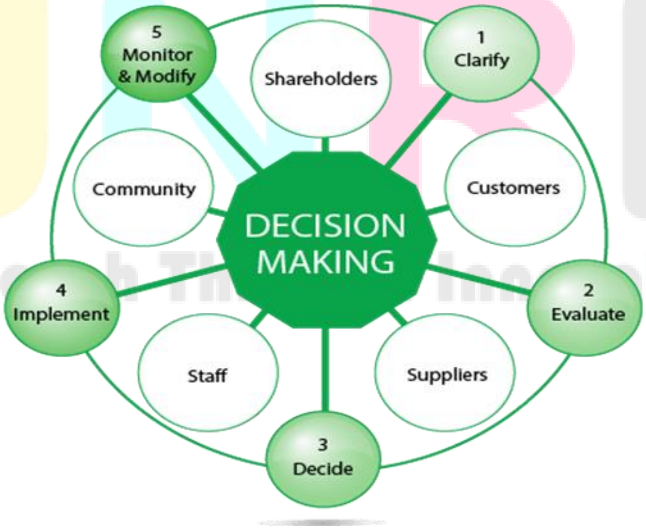
- The situational awareness of autonomous systems is enhanced by XAI. Operators and users can comprehend how the system perceives its surroundings and reacts to different circumstances when there are clear explanations provided.

7. Adaptability and Learning:

- Experiences are a common source of learning for autonomous systems. Through XAI, users may understand how the system updates and adjusts its knowledge, ensuring that learning procedures match the intended goals.

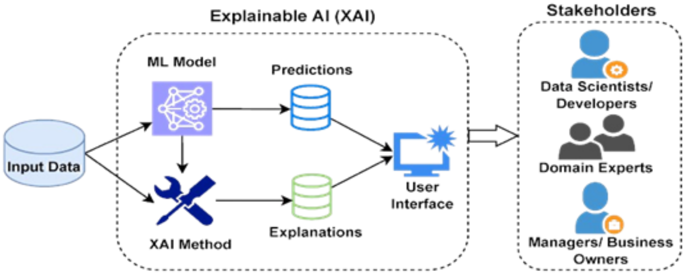
8. Ethical Decision-Making:

- XAI supports ethical considerations in autonomous systems by providing visibility into the ethical frameworks and principles embedded in decision-making algorithms. This is crucial for ensuring responsible and morally sound behavior.



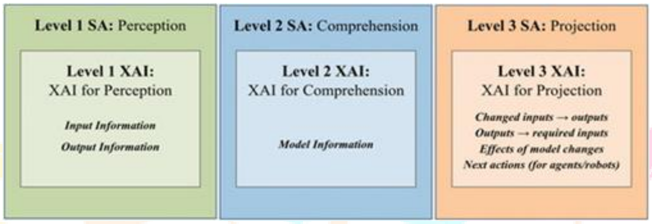
9. Real-time Feedback:

- Giving explanations in real time enables quick feedback on the operations of the system. In dynamic and uncertain situations, users can intervene or steer the system with the help of this feedback loop.
- Real time artificial intelligence is a subfield of AI that deals with the processing of information as it comes in.



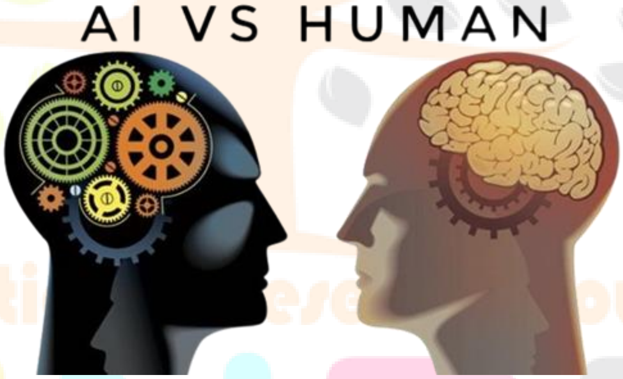
10. Explanations for Control Actions:

• Critical decisions about control actions are frequently made by autonomous systems. XAI offers justifications for these choices, assisting users in comprehending the rationale behind a specific action and enabling informed modifications as needed.



11. Human Oversight and Intervention:

• Effective supervision of autonomous systems by human operators is made possible by XAI. Being able to comprehend the reasoning behind decisions enables humans to step in when necessary, especially when the system's actions could have serious repercussions.



12. Training and Education:

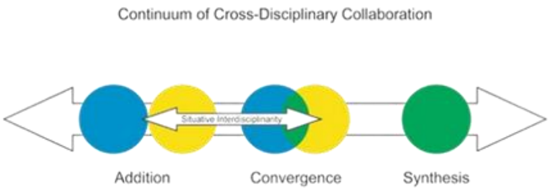
• XAI facilitates training and education programs for operators and stakeholders. Transparent explanations aid in conveying the system's capabilities, limitations, and operational characteristics.

13. Public Acceptance:

• Transparency through XAI contributes to public acceptance of autonomous systems. Providing clear explanations helps mitigate concerns, fears, and misconceptions surrounding the deployment of AI-driven autonomous technologies.

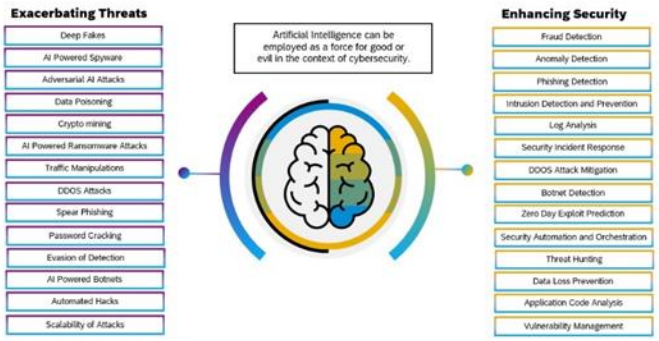
14. Cross-disciplinary Collaboration:

• Collaboration between domain experts, engineers, and AI professionals is encouraged by XAI. Comprehending the demands of diverse stakeholders regarding interpretability guarantees that self-governing systems fulfill both technical and functional prerequisites.


Continuum of Cross-Disciplinary Collaboration
Situative Interdisciplinarity
Addition
Convergence
Synthesis

15. Security Considerations:

- XAI helps by enabling users to closely examine the decision-making process, which helps to address security problems. This aids in locating weak points and possible dangers related to hostile assaults on independent systems.



Privacy-Preserving XAI:

- Privacy concerns arise when dealing with sensitive data in XAI. Security measures, such as encryption and secure computation, should be in place to protect the privacy of individuals and prevent unauthorized access to confidential information during explanation generation.

Secure Data Handling:

- To mitigate the risk of data breaches, XAI systems should incorporate secure data handling practices. This includes encrypting data, implementing strict access controls, and ensuring secure transmission of data during both training and operational phases.

Adversarial Attacks:

- Explanation mechanisms in XAI models may be susceptible to adversarial attacks where attackers manipulate input data to deceive the system. Security measures must be in place to identify and mitigate these attacks without compromising the interpretability of the model.

XAI Program Development and Progress

Figure 3. XAI Research Teams 1 (TA1) developer teams and the TA2 team [from the Florida Institute for Human and Machine Cognition (IHMC)] that is developing the psychologic model of explanation. Autonomy and data analytics are the two challenge problem areas that three TA1 teams are pursuing; three are focusing solely on the former, and five are focusing only on the latter. The TA1 teams are looking into a wide range of methods for creating explainable models and explanation interfaces, in accordance with the tactics shown in figure 2.

Naturalistic Decision-Making Foundations of XAI

The IHMC team, comprising researchers from Macro Cognition and Michigan Technological University, aims to create and assess psychologically realistic models of explanation. Their goal is to formulate practical concepts, methods, measures, and metrics for explanatory reasoning. The IHMC team is actively engaged in the development and evaluation of these models.

The IHMC team is delving into the essence of explanation, exploring the conditions that lead to a person's satisfaction with explanations regarding both the workings of intricate systems and the reasons behind their actions in specific situations. To tackle these inquiries, the team has devised a naturalistic model of human explanatory reasoning. They are offering guidance to performer teams on evaluating the effectiveness of their eXplainable Artificial Intelligence (XAI) systems' explanations. Drawing from philosophy of science and psychology literature, the team has synthesized criteria for assessing the quality of explanations.

Additionally, the team is systematically collecting and analyzing cases where individuals create or receive explanations for complex systems. They have developed measures for evaluating explanation quality, a user's mental model (including correctness and completeness), and user task performance. These measures enable users to make informed judgments about when to trust or doubt the system, particularly by exploring the decision-making processes and performance of XAI systems, including considerations for boundary cases and potential vulnerabilities such as spoofing deep neural networks (DNNs). This methodology has been detailed in a series of essays authored by Hoffman and Klein. Figure 8 illustrates IHMC's model of the XAI explanation process, outlining measurement categories for assessing explanation effectiveness. In this model, the user receives a recommendation or decision from an XAI system, accompanied by an explanation that can be tested for quality against predefined criteria and user satisfaction. The explanation contributes to the user's mental model of the AI system, which is subject to testing for accuracy and comprehension. The user's trust in the AI system, as well as their task performance, is influenced by the AI system's recommendations and the user's mental model, both of which can be measured. This comprehensive approach aids the XAI evaluator in testing the developer teams' XAI systems.

Evaluation

The XAI program's independent government evaluator is the Naval Research Laboratory. For Phase 1, the laboratory (with IHMC's help) prepared an evaluation framework for the TA1 teams to use as a template for

The IHMC team is in the process of designing and executing their Phase 1 evaluation experiments. In this phase, they will choose a test problem or problems within the challenge areas of data analytics or autonomy. They plan to apply their newly developed machine learning (ML) techniques to learn an interpretable model for these problems. The evaluation process involves assessing the performance of the ML model, as outlined in Table 1. Subsequently, the team intends to integrate their learned model with an explanation interface to create an explainable learning system.

The experimentation will entail users performing specific tasks using the explainable learning system. The effectiveness of explanations will be measured according to IHMC's model of the explanation process (depicted in Figure 4) and the explanation effectiveness measurement categories (detailed in Table 1).

The evaluation will encompass several experimental conditions:

1. Without Explanation:
   - The XAI system is used to execute a task without providing explanations to the user.

2. With Explanation:
   - The XAI system is employed to carry out a task and generates explanations for each recommendation, decision, and action it performs.
- The XAI system is utilized for task execution, generating only partial or ablated explanations to assess various features of explanations.

4. Control:
   - A baseline state-of-the-art nonexplainable system is employed to execute a task.

Through these experimental conditions, the IHMC team aims to comprehensively evaluate the performance and effectiveness of their explainable learning system, providing insights into the impact of explanations on user interaction and task outcomes.

Deeply Explainable AI

Researchers from Boston University, the University of Amsterdam, and Kitwara are part of the University of California, Berkeley (UCB) team that is creating an artificial intelligence (AI) system that is understandable by humans because explicit structural interpretation (Hu et al. 2017), provides post hoc (Park et al. 2018) and introspective (Rama Nishka et al. 2017) explanations, has predictive behavior, and allows for appropriate trust (Huang et al. 2018). Selecting the most meaningful explanations for a user and producing accurate explanations of model behavior are the main challenges facing deeply explainable AI (DEXAI). In order to address the former, UCB is developing explicit or implicit explanation models. Specifically, they are able to construct naturally understandable explicit structures or implicitly present complex latent representations in ways that are easily understood. These DEXAI models generate a range of potential explanation strategies. These actions are referred to as reflexive because they are generated in the absence of any user model. UCB is addressing both challenge problem areas. For autonomy, DEXAI will be demonstrated in vehicle control (using the Berkeley Deep Drive data set and the CARLA simulator) (Kim and Canny 2017) and strategy game scenarios (StarCraft II). For data analytics, DEXAI will be demonstrated using visual question answering (VQA) and filtering tasks (for example, using large-scale data sets such as VQA-X and ACT-X for VQA tasks and activity recognition tasks, respectively), xView, and Distin.

XAI Toolkits and Frameworks:

Explainable Artificial Intelligence (XAI) toolkits and frameworks are essential components for researchers and developers working on building AI systems with interpretability and transparency. These tools assist in creating, evaluating, and deploying models that provide understandable explanations for their decisions. Here are some notable XAI toolkits and frameworks:

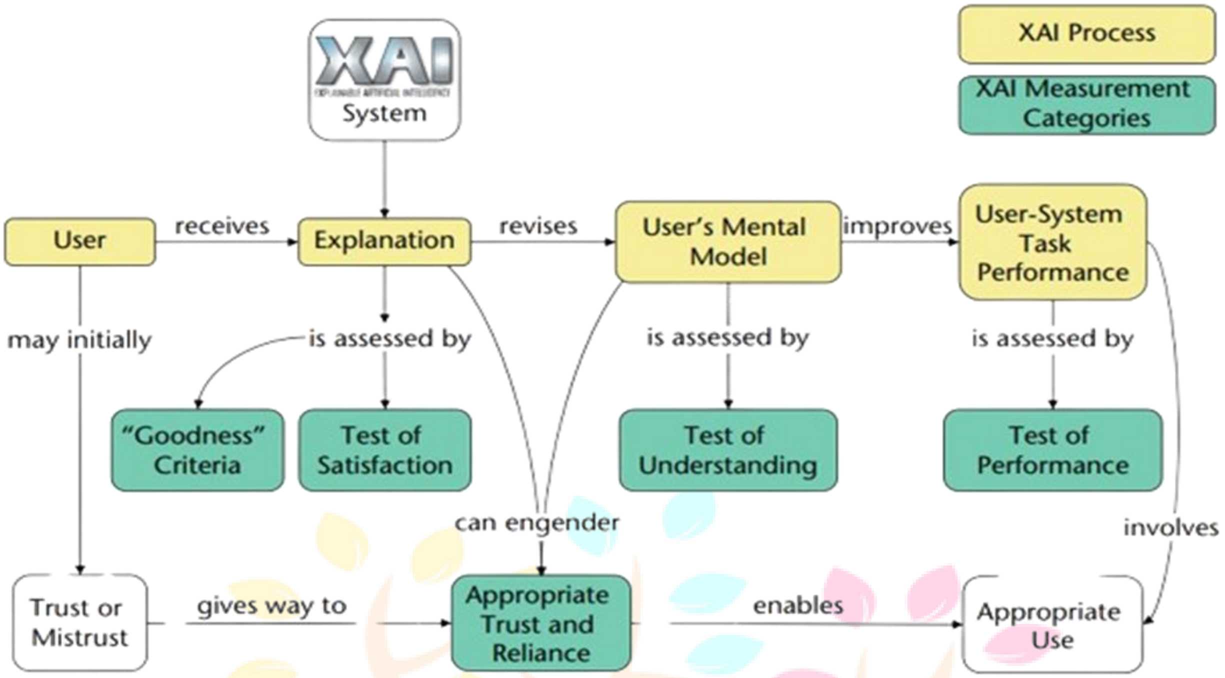
1. SHAP (SHapley Additive exPlanations):
   - Overview: SHAP is a popular Python library that implements Shapley values for explaining the output of any machine learning model.
   - Key Features:
     - Provides both global and local interpretability.
     - Offers consistent and theoretically grounded explanations.

2.LIME (Local Interpretable Model-agnostic Explanations):
   - Overview:

LIME is a framework for explaining the predictions of machine learning models by approximating them with locally interpretable models.
   - Key Features: - Model-agnostic approach allows it to be used with various types of models.
     - Focuses on generating locally faithful explanations for individual predictions.

3. Interpret:
   - Overview: InterpretML is an open-source Python library that aims to simplify the interpretation of machine learning models.
   - Key Features:
     - Provides a unified interface for multiple interpretability techniques.
     - Supports a variety of model-agnostic and model-specific interpretability methods.

4. AIX360 (AI Explainability 360):
   - Overview: AIX360 is an open-source toolkit developed by IBM that provides a comprehensive set of algorithms for addressing different facets of explainability.
   - Key Features:
     - Offers a diverse range of algorithms, from rule-based methods to post-hoc approaches.



5. ELI5 (Explain Like I'm 5):
- Overview: ELI5 is a Python library that allows you to visualize and debug various machine learning models.
- Key Features:
  - Supports a wide range of models, including scikit-learn, XGBoost, and more.
  - Provides feature importances, permutation importance, and other explanation methods.

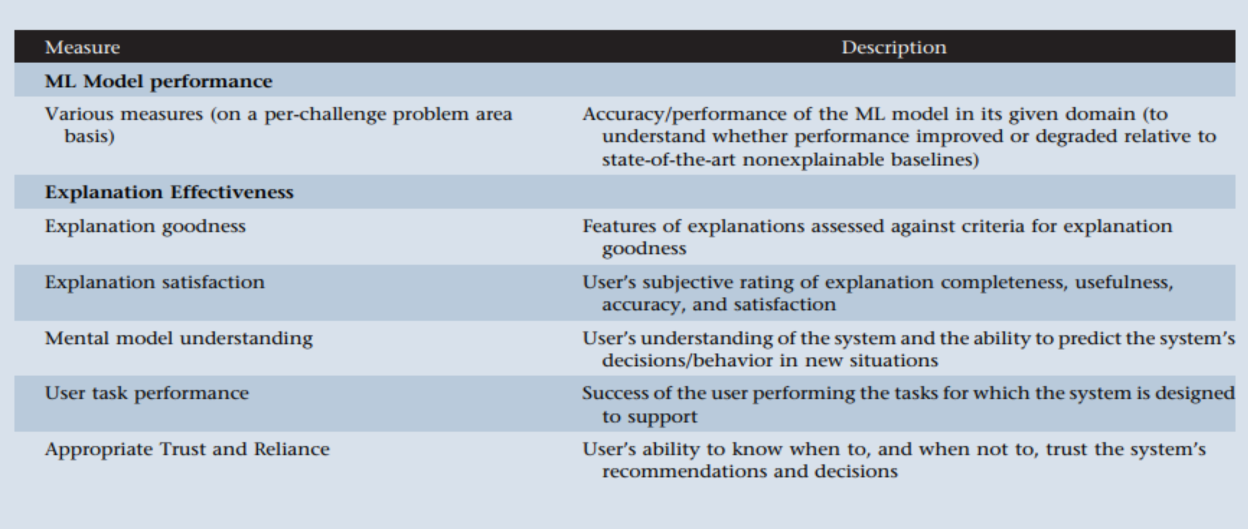
6. DALEX (Descriptive mAchine Learning EXplanations):
- Overview: DALEX is an R package designed for model-agnostic exploration, explanation, and validation of complex machine learning models.
- Key Features:
  - Enables the visualization of complex relationships in a simple and understandable way.
  - Facilitates the comparison of model performance and predictions.

7. TensorFlow Lattice:
- Overview: TensorFlow Lattice is an open-source library that integrates with TensorFlow for building interpretable machine learning models using lattice functions.

- Key Features:
  - Focuses on building models with monotonic relationships, enhancing interpretability.
  - Suitable for tasks where understanding the input-output relationships is crucial.

8. SKATER (Model Agnostic Insights and Diagnostics for Machine Learning):
- Overview: SKATER is a Python library that provides a set of tools for model-agnostic interpretation, diagnostics, and debugging of machine learning models.
- Key Features:
  - Supports rule-based and tree-based explanations.
  - Offers visualization tools for better understanding model behavior.

When choosing an XAI toolkit or framework, it's essential to consider the specific requirements of your project, the type of model you are working with, and the interpretability techniques that align with your goals. Additionally, staying informed about updates and new tools in the rapidly evolving field of XAI.



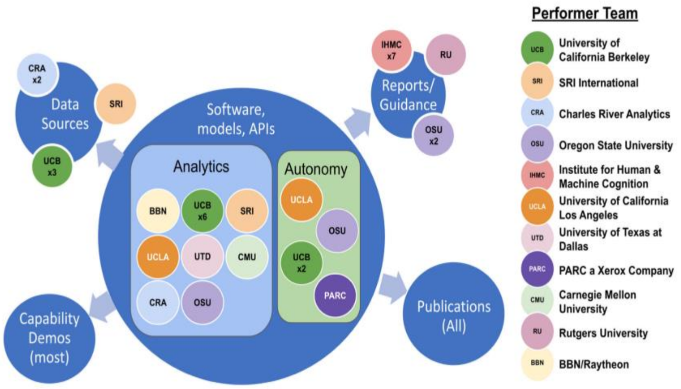
Measure	Description
ML Model performance	
Various measures (on a per-challenge problem area basis)	Accuracy/performance of the ML model in its given domain (to understand whether performance improved or degraded relative to state-of-the-art nonexplainable baselines)
Explanation Effectiveness	
Explanation goodness	Features of explanations assessed against criteria for explanation goodness
Explanation satisfaction	User's subjective rating of explanation completeness, usefulness, accuracy, and satisfaction
Mental model understanding	User's understanding of the system and the ability to predict the system's decisions/behavior in new situations
User task performance	Success of the user performing the tasks for which the system is designed to support
Appropriate Trust and Reliance	User's ability to know when to, and when not to, trust the system's recommendations and decisions

Table 1. Measurement Categories.

TOOLKIT COMPONENTS

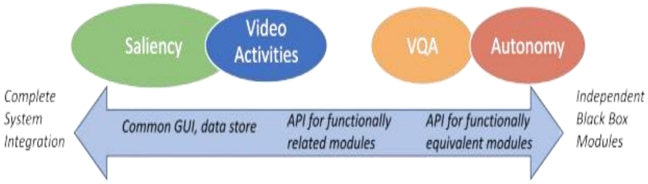
Non-software artifacts: - As discussed above, the toolkit will explicitly include non-software components. Each of these artifacts will be kept in an extensive, user-accessible, central repository that connects related objects. This repository will be accessible

Components of the XAI Toolkit address a spectrum of program artifacts. Non-software artifacts include publications, reports and guidance, and data sources. Software artifacts include demonstrations of various XAI capabilities in addition to domain-specific software frameworks. The diagram shows the expected toolkit contributions to the appropriate organizational structure for the XAI Toolkit Working Group (XTWG) from the various DARPA XAI performer teams.

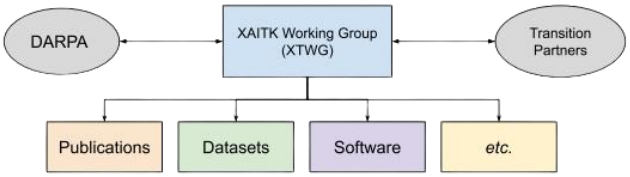

Data Sources
CRA x2
SRI
UCB x3
Software, models, APIs
Analytics
BBN
UCB x6
SRI
UCLA
UTD
CMU
CRA
OSU
Autonomy
UCLA
OSU
UCB x2
PARC
Reports/ Guidance
IHMC x7
RU
OSU x2
Capability Demos (most)
Publications (All)
Performer Team
UCB University of California Berkeley
SRI SRI International
CRA Charles River Analytics
OSU Oregon State University
IHMC Institute for Human & Machine Cognition
UCLA University of California Los Angeles
UTD University of Texas at Dallas
PARC PARC a Xerox Company
CMU Carnegie Mellon University
RU Rutgers University
BBN BBN/Raytheon

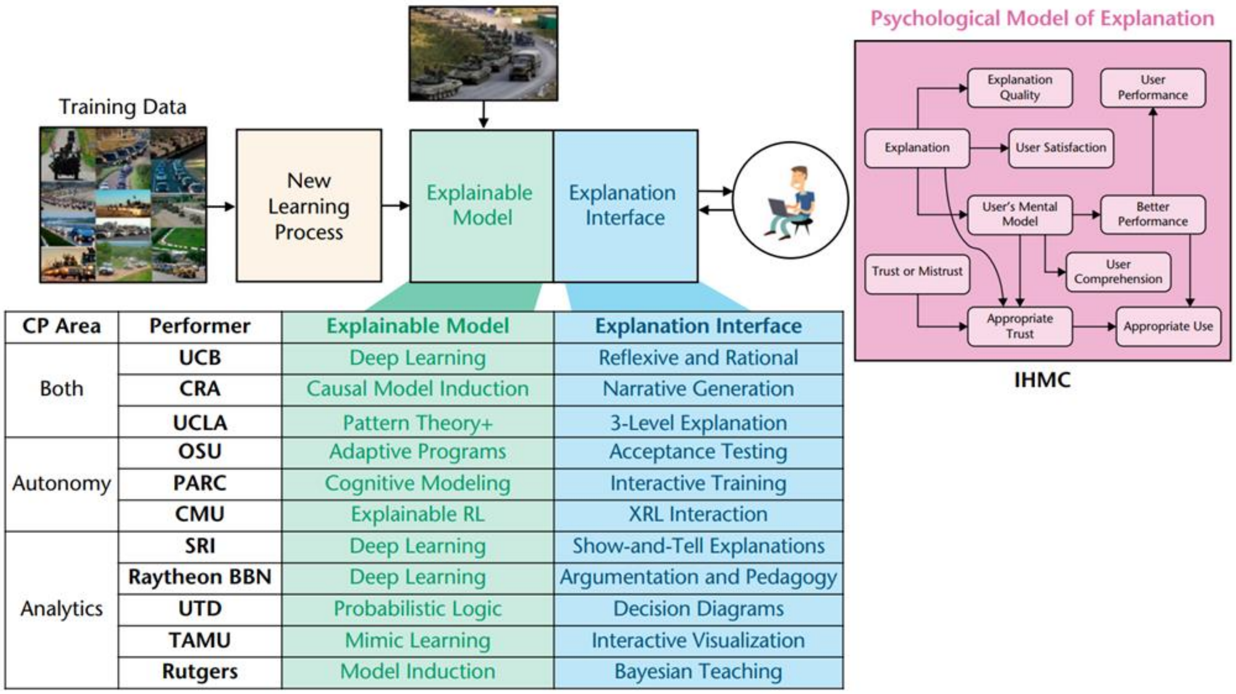
The XTWG (blue box) is notionally responsible for collecting and curating the different artifacts (e.g., publications, datasets, software, etc.) from the DARPA XAI program (bottom row). Additionally, the XTWG serves as a liaison between DARPA and possible transition partners, or "gray ovals," locating pertinent use cases for XAI that can guide the creation of toolkits.

Degrees of software integration. The XAI toolkit is expected to receive contributions in many forms, from standalone modules to fully integrated systems. Based on an initial assessment of these contributions, we have grouped together certain contributions and placed them along the software integration spectrum. For instance, there are numerous planned contributions related to saliency, which enables the creation of a common software framework.


Saliency
Video Activities
VQA
Autonomy
Complete System Integration
Common GUI, data store
API for functionally related modules
API for functionally equivalent modules
Independent Black Box Modules

to users via a public-facing website: https://xaitk.org We expect that periodic audits will be necessary to make sure these assets are kept up to date during their subsequent management.





CP Area	Performer	Explainable Model	Explanation Interface
Both	UCB	Deep Learning	Reflexive and Rational
	CRA	Causal Model Induction	Narrative Generation
	UCLA	Pattern Theory+	3-Level Explanation
Autonomy	OSU	Adaptive Programs	Acceptance Testing
	PARC	Cognitive Modeling	Interactive Training
	CMU	Explainable RL	XRL Interaction
Analytics	SRI	Deep Learning	Show-and-Tell Explanations
	Raytheon BBN	Deep Learning	Argumentation and Pedagogy
	UTD	Probabilistic Logic	Decision Diagrams
	TAMU	Mimic Learning	Interactive Visualization
	Rutgers	Model Induction	Bayesian Teaching

Figure 3. XAI Research Teams

Autonomy Domain Implementation:

Deep reinforcement learning aims to acquire adaptable action policies capable of governing autonomous systems across diverse environments. Numerous eXplainable Artificial Intelligence (XAI) methods have been created to elucidate the actions and decisions made by these deep reinforcement learning systems. Figure 7 illustrates a conceptual diagram of a hypothetical autonomy Domain Implementation, facilitating the investigation of pivotal state algorithms for reinforcement learning. In this context, it is assumed that the client application is tailored to a specific task, such as a video game or a self-driving vehicle. The framework allows for a comparison of various algorithms within this specialized application domain.

TOOLKIT TRANSITION EFFORT

Concurrently with the development of the toolkit, we expect that efforts towards transitioning will involve the identification and outreach to potential collaborators. These interactions, as outlined below, are foreseen to have a significant impact on the design and implementation of the eXplainable Artificial Intelligence Toolkit (XAITK) as it progresses. Close collaboration with interested partners is crucial for identifying domains, The datasets, and workflows that could benefit from eXplainable AI (XAI) and the XAITK. The development of the toolkit will be guided by these identified use cases, ensuring that the research code produced by the DARPA XAI program can be effectively integrated into deployed systems.

Figure 8 illustrates potential use cases of XAI with different points of integration into the system. Explanations can be incorporated into the underlying data as a form of preprocessing, contribute to building interpretable models within the system, or be presented to users in a post-hoc manner, such as through methods like saliency maps. Various combinations of these approaches may also be employed, and understanding the level at which XAI is utilized is crucial for comprehending potential use cases and the required level of system integration.

The transition effort is expected to encompass outreach and education, development, dataset integration, codebase integration, and evaluation on partner data. Successful transition of the XAITK involves identifying suitable partners, engaging with them to communicate current and near-term XAI capabilities, and exploring opportunities where XAI could add value. After identifying interested partners, an in-depth analysis of their workflows will be conducted to pinpoint specific scenarios where XAI can provide assistance. The team will estimate the work required versus the expected benefits, leading to a decision on whether to proceed further. If interest persists, the XAITK will collaborate closely with the partner's existing data and software infrastructure. The partner will contribute relevant data, aiding in the assessment of XAITK performance, model training, and highlighting software engineering considerations necessary for integration (such as file formats and metadata). Extensive discussions on software integration pathways and requirements will continue, especially when working with non-public or restricted data.
XAITK team will also identify potential avenues for demonstrations and evaluations to disseminate the toolkit's capabilities widely, regardless of the integration level. As a final objective, formalized evaluations of XAITK capabilities will be pursued on partner-supplied data, either qualitatively with ground-truth or quantitatively using subject matter expert/analyst assessments, questionnaires, etc. Throughout these efforts, the XAI Transition Working Group (XTWG) will play an active role as a liaison between transition partners and DARPA.

CONCLUSION;-

The field of eXplainable Artificial Intelligence (XAI) holds immense promise for enhancing the trust, transparency, and adoption of AI systems across various domains. As we conclude our exploration of XAI, several key takeaways emerge:

1. Enhancing Trust and Adoption:
- XAI addresses one of the significant challenges in deploying AI systems – the lack of transparency. By providing interpretable explanations for AI decisions, XAI enhances trust among users, stakeholders, and the general public. This, in turn, promotes wider acceptance and adoption of AI technologies.

2. Interpretable Models for Real-World Impact:
- The development and integration of interpretable models are crucial for the real-world impact of AI applications. XAI techniques enable practitioners to understand and validate the decision-making processes of complex models, fostering accountability and reliability in critical applications such as healthcare, finance, and autonomous systems.

3. Human-Centric AI Development:
- XAI emphasizes the importance of a human-centric approach to AI development. Understanding how users perceive and interact with AI systems is essential for designing explanations that are not only accurate but also meaningful and user-friendly. This human-in-the-loop paradigm contributes to the responsible and ethical deployment of AI.

4. Domain-Specific Adaptability:
- XAI recognizes the need for adaptability in different domains. No one-size-fits-all solution exists, and XAI techniques must be tailored to specific applications and user requirements. This adaptability ensures that explanations are relevant and useful in diverse contexts, ranging from medical diagnoses to financial predictions.

5. Addressing Bias and Fairness:
- XAI plays a pivotal role in addressing issues of bias and fairness in AI systems. By making the decision-making processes transparent, XAI facilitates the identification and mitigation of biases, contributing to the development of fair and equitable AI solutions.

6. Continuous Collaboration and Research:
- The evolution of XAI is an ongoing process that requires continuous collaboration between researchers, practitioners, and policymakers. As AI technologies advance, new challenges and opportunities arise, necessitating innovative XAI techniques. Ongoing research and collaboration are essential to keep pace with the dynamic landscape of AI.

In conclusion, XAI stands as a cornerstone for the responsible and ethical development of AI systems. Its ability to demystify complex models, empower users, and ensure fairness positions XAI as a critical component in shaping the future of artificial intelligence. As we move forward, the commitment to transparency, user-centric design, and interdisciplinary collaboration will be paramount in harnessing the full potential of XAI for the benefit of society.

The main contribution of this study is a taxonomy of the existing machine learning interpretability methods that allows for a multi-perspective comparison among them. Under this taxonomy, four major categories for interpretability methods were identified: methods for explaining complex black-box models, methods for creating white-box models, methods that promote fairness and restrict the presence of discrimination, and, finally, techniques for evaluating the model predictions' sensitivity. Due to the research community's strong focus on deep learning, neural networks have dominated the majority of the literature on interpretability techniques. and their applications to computer vision and natural language processing. The majority of interpretability techniques for elucidating deep learning models make use of image classification and generate saliency maps that illustrate the relative importance of various image regions. In many cases, this is achieved through exploiting the gradient information flowing through the layers of the network, Grad-CAM, a direct extension of , being a prime and most influential example in terms of citations per year Using deconvolutional neural networks is another method (the most influential overall) for producing saliency maps using the same metric. In terms of explaining any black-box model, the LIMEand SHA methods are, by far, the most comprehensive and dominant across the literature methods for visualizing feature interactions and feature importance, while Friedman's PDPs, although much Though less sophisticated and older, it's still a well-liked option. In addition to being model-neutral, the LIME and SHAP approaches have been shown to work with any kind of data. It is very difficult to develop white-box high performance models, particularly in computer vision and natural language processing where there is an unbridgeable performance gap with deep learning models. Furthermore, white-box models, which are limited to performing well in a single task, are losing traction in the literature and are rapidly losing interest as models are expected more than ever to be competitive on multiple tasks and knowledge transfer from one domain to another is becoming a common theme. Caruana et al.'s work is the most noteworthy in this category. who proposed a version of generalized additive models with pairwise interactions (GA2Ms), originally proposed in , reporting state-of-the-art accuracy in two healthcare applications.

--***--

~Authors:

Naitik Ashok Pawar
https://www.linkedin.com/in/naitik-pawar-784957253

saikumar Madel
https://www.linkedin.com/in/saikumar-madel-9b318925a

Sanika mukhmale
https://www.linkedin.com/in/sanika-mukhmale-a860642b1

Mugdha mule
https://in.linkedin.com/in/mugdha-mule-1b4467299

~Funding: This research received no external funding.

~Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Repository Links

Tool	Repository Link
Grad-CAM	https://github.com/naitikpawar22/grad-cam
accessorize-to-a-crime	https://github.com/naitikpawar22/accessorize-to-a-crime.git
adversarial-squad	https://github.com/naitikpawar22/adversarial-squad.git
adversarial_text	https://github.com/naitikpawar22/adversarial_text.git
AIX360	https://github.com/naitikpawar22/AIX360
Aequitas	https://github.com/naitikpawar22/aequitas
adversar	https://github.com/naitikpawar22/adversarial_training_methods
Adversarial_training	https://github.com/naitikpawar22/adversarial_training
fair-classification	https://github.com/naitikpawar22/fair-classification
boundary-attack	https://github.com/naitikpawar22/boundary-attack
equalized_odds_and_calibration	https://github.com/naitikpawar22/equalized_odds_and_calibration
Eli5	https://github.com/naitikpawar22/eli5
DLIME	https://github.com/naitikpawar22/dlime_experiments
DeepLift	https://github.com/naitikpawar22/deeplift
DeepExplain	https://github.com/naitikpawar22/DeepExplain
Deep Visualization Toolbox	https://github.com/naitikpawar22/deep-visualization-toolbox
Alibi	https://github.com/naitikpawar22/alibi
CAM	https://github.com/naitikpawar22/CAM
AnalysisBySynthesis	https://github.com/naitikpawar22/AnalysisBySynthesis
Fairness	https://github.com/naitikpawar22/fairness
FairMachineLearning	https://github.com/naitikpawar22/FairMachineLearning
Fairlearn	https://github.com/naitikpawar22/fairlearn
fair-classification	https://github.com/naitikpawar22/fair-classification

REFERENCES

1. Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable AI systems for the medical domain? arXiv: 1712.09923; 2017.
2. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv:1702.08608; 2017.
3. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Mach Intell. 2019;1(5):206-215.
4. Samek W, Wiegand T, Müller KR. Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. ITU Journal: ICT Discoveries, 2017(1).
5. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access. 2018;6:52138-52160.
6. Arrieta AB, Daíz-Rodríguez N, Del Ser J, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion. 2020;58:82-115.
7. Vilone G, Longo L. Explainable artificial intelligence: a systematic review. arXiv:2006.00093; 2020.
8. Board DI. AI principles: recommendations on the ethical use of artificial intelligence by the Department of Defense. Supporting document, Defense Innovation Board 2019.
9. Belbute-Peres, F., and Kolter, J. Z. 2017. A Modular Differentiable Rigid Body Physics Engine. Paper presented at the Neural Information Processing Systems Deep Reinforcement Learning Symposium. Long Beach, CA, December 7
10. Chakraborty, S.; Tomsett, R.; Raghavendra, R.; Harborne, D.; Alzantot, M.; Cerutti, F.; and Srivastava, M.; et al. 2017. Interpretability of Deep Learning Models: A Survey of Results. Presented at the IEEE Smart World Congress 2017 Workshop: DAIS 2017 — Workshop on Distributed Analytics Infrastructure and Algorithms for Multi-Organization Federations, San Francisco, CA, August 4–8. doi.org/10. 1109/UIC-ATC.2017.8397411
11.J. Zhu et al.
Explainable AI for designers: A human-centered perspective on mixed-initiative co-creation
2018 IEEE Conference on Computational Intelligence and Games (CIG)
(2018)
12.T. Miller
Explanation in artificial intelligence: Insights from the social sciences
Artif. Intell.
(2019)
13.G. Montavon et al.
Methods for interpreting and understanding deep neural networks
Digital Signal Processing
(2018)
14.R.S. Michalski
A theory and methodology of inductive learning
Machine learning
(1983)
15.D. Martens et al.
Performance of classification models from a user perspective
Decision Support Systems
(2011)
16.G. Montavon et al.
Explaining nonlinear classification decisions with deep taylor decomposition
Pattern Recognition
(2017)
17.J.D. Olden et al.
Illuminating the “black box”: a randomization approach for understanding variable contributions in artificial neural networks
Ecological modelling
(2002)
18.B. Üstün et al.
Visualisation and interpretation of support vector regression models
Analytica Chimica Acta
(2007)
19.J. Huysmans et al.
An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models
Decision Support Systems
(2011)
20.D. Martens et al.
Comprehensible credit scoring models using rule extraction from support vector machines
European Journal of Operational Research

(2007)
21.R. Krishnan et al.
Extracting decision trees from trained neural networks
Pattern Recognition
(1999)
22.S.J. Russell et al.
Artificial intelligence: a modern approach
(2016)
23.D.M. West
The future of work: robots, AI, and automation
(2018)
24.B. Goodman et al.
European union regulations on algorithmic decision-making and a “right to explanation”
AI Magazine
(2017)
25.D. Castelvecchi
Can we open the black box of AI?
Nature News
(2016)
26.Z.C. Lipton
The mythos of model interpretability
Queue
(2018)
27.A. Preece, D. Harborne, D. Braines, R. Tomsett, S. Chakraborty, Stakeholders in Explainable AI,...
28.D. Gunning
Explainable artificial intelligence (xAI)
Technical Report
(2017)
29.E. Tjoa, C. Guan, A survey on explainable artificial intelligence (XAI): Towards medical XAI,...
30.F.K. Došilović et al.
Explainable artificial intelligence: A survey
41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)
31.P. Hall, On the Art and Science of Machine Learning Explanations,...
32.L.H. Gilpin, D. Bau, B.Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining Explanations: An Overview of...

33..Interpretable Machine Learning Journal (IMLJ): A journal dedicated to research on interpretable machine learning and XAI.
34.Journal of Artificial Intelligence Research (JAIR): Covers a wide range of AI topics, including research on explainability and interpretability.
35.Transactions on Knowledge and Data Engineering (TKDE): Publishes research on data and knowledge engineering, including aspects related to XAI.
36. Interpretable Machine Learning Journal (IMLJ):This journal is specifically dedicated to research on interpretable machine learning and XAI.
37. AI and Ethics:This journal often publishes research related to the ethical implications of AI, including discussions on transparency and fairness.
38. Journal of Artificial Intelligence Research (JAIR): JAIR covers a wide range of AI topics, and it often includes papers related to explainability, interpretability, and transparency in AI systems.
39. Nature Machine Intelligence: A multidisciplinary journal that publishes research in artificial intelligence, including topics related to XAI.
40. Journal of Machine Learning Research (JMLR): JMLR publishes research in machine learning, and some of its articles may touch upon interpretability and explainability.
41. ACM Transactions on Interactive Intelligent Systems (TiiS): This ACM journal focuses on research at the intersection of artificial intelligence and human-computer interaction, including XAI.
42. Data Mining and Knowledge Discovery:This journal often includes research on interpretable machine learning and the extraction of knowledge from complex data.
43. Journal of Artificial Intelligence Research (JAIR): A leading journal covering all areas of AI research, including machine learning, knowledge representation, reasoning, planning, and more.
44.. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI): Focuses on pattern analysis and machine intelligence, covering a wide range of topics within computer vision and machine learning.
45. Machine Learning: A journal that publishes articles on a wide range of topics related to machine learning, including theoretical aspects and practical applications.
46. ACM Transactions on Intelligent Systems and Technology (TIST): An ACM journal that focuses on the integration of AI technologies, including intelligent systems and advanced technologies.
47. Science Robotics:While not exclusively focused on AI, this journal covers a broad spectrum of robotics research, including AI-driven advancements.

48.Computer Vision and Image Understanding (CVIU):Focuses on computer vision research, including image understanding, pattern recognition, and related AI topics.
49. IEEE Transactions on Artificial Intelligence (T-AI): Publishes high-quality research articles and reviews covering various aspects of artificial intelligence.
50. Jin, D.; Jin, Z.; Zhou, J.T.; Szolovits, P. Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, the Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, the Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, 7–12 February 2020; AAAI Press: Palo Alto, CA, USA, 2020; pp. 8018–8025.
51. Garg, S.; Ramakrishnan, G. BAE: BERT-based Adversarial Examples for Text Classification. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, 16–20 November 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 6174–6181.
52. Li, L.; Ma, R.; Guo, Q.; Xue, X.; Qiu, X. BERT-ATTACK: Adversarial Attack Against BERT Using BERT. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, 16–20 November 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 6193–6202.
53. Tan, S.; Joty, S.; Kan, M.Y.; Socher, R. It's Morphin' Time! Combating Linguistic Discrimination with Inflectional Perturbations. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 2920–2935. [CrossRef]
54. Zang, Y.; Qi, F.; Yang, C.; Liu, Z.; Zhang, M.; Liu, Q.; Sun, M. Word-level textual adversarial attacking as combinatorial optimization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 6066–6080

ere;DONOTplacethemonthefirstpageofyourpaperorasafootnote.

REFERENCES

[1] Ali, A. 2001.Macroeconomic variables as common pervasive risk factors and the empirical content of the Arbitrage Pricing Theory. Journal of Empirical finance, 5(3): 221–240.
[2] Basu, S. 1997. The Investment Performance of Common Stocks in Relation to their Price to Earnings Ratio: A Test of the Efficient Markets Hypothesis. Journal of Finance, 33(3): 663-682.
[3] Bhatti, U. and Hanif. M. 2010. Validity of Capital Assets Pricing Model.Evidence from KSE-Pakistan.European Journal of Economics, Finance and Administrative Science, 3 (20).