Introduction to Explainable AI

Chapter · July 2024

DOI: 10.1007/978-981-97-3705-5_1

8 authors, including:

Amit Ganatra
Parul University
183 PUBLICATIONS **2,434** CITATIONS
SEE PROFILE

Brijeshkumar Y. Panchal
Sardar Vallabhbhai Patel Institute of Technology
38 PUBLICATIONS **100** CITATIONS
SEE PROFILE

Devarshi Doshi
Charotar University of Science and Technology
1 PUBLICATION **4** CITATIONS
SEE PROFILE

Bijal Talati
sardar vallabhbhai patel institute of technology,vasad
7 PUBLICATIONS **16** CITATIONS
SEE PROFILE

Introduction to Explainable AI

Amit Ganatra, Brijeshkumar Y. Panchal, Devarshi Doshi, Devanshi Bhatt, Jesal Desai, Bijal Talati, Neha Soni, and Apurva Shah

Abstract Explainable AI (XAI) has emerged as an essential realm aimed at tackling the opacity of intricate AI models and nurturing confidence in their judgments. This study extensively investigates the fundamental underpinnings, methodologies, and practical implementations of XAI. The bedrock of XAI lies in the urgency to demystify the internal mechanisms of AI models, rendering their decision-making transparent for human stakeholders. Within the domain of XAI, diverse methodologies encompass a spectrum of techniques such as interpretable models, scrutiny of feature significance, localized and holistic elucidations, visual representations, and explications in natural language. These methodologies collectively foster intelligibility and amplify the explicable nature of AI models. This research significantly enriches the expanding reservoir of scholarly exploration by clarifying the core tenets of XAI. This comprehensive survey unmistakably demonstrates that XAI assumes a pivotal role in bridging the chasm between intricate AI processes and human comprehension. Consequently, it clears the path for a more reliable and efficacious partnership between human intellect and mechanical ingenuity.

A. Ganatra
Parul University, Limda, Waghodia, Vadodara, Gujarat, India

B. Y. Panchal (✉) · N. Soni
Computer Engineering Department, Sardar Vallabhbhai Patel Institute of Technology (SVIT), Vasad, Anand, Gujarat Technological University (GTU), Ahmedabad, Gujarat, India

D. Doshi · D. Bhatt · J. Desai
Department of Computer Science and Engineering, Devang Patel Institute of Advance Technology and Research (DEPSTAR), Faculty of Technology and Engineering (FTE), Charotar University of Science and Technology (CHARUSAT), Anand, India

B. Talati
Department of Computer Science and Engineering, Parul Institute of Technology [PIT], Parul University, Limda, Waghodia, Vadodara, Gujarat, India

A. Shah
Computer Science and Engineering Department, Faculty of Technology and Engineering, The Maharaja Sayajirao University of Baroda, Vadodara, Gujarat, India

R. Aluvalu et al. (eds.), *Explainable AI in Health Informatics*, Computational Intelligence Methods and Applications,
https://doi.org/10.1007/978-981-97-3705-5_1

Keywords Explainable AI · Artificial intelligence · Deep learning · XAI · Healthcare · Automation

1 Introduction

AI stands for Artificial Intelligence. It is a division of computer science that helps in developing or designing a system that can show an intelligent behavior. At the core level, AI is the science and engineering of making intelligent machines, particularly computer programs or systems. In recent years, an AI technique has been successfully engaged to solve a wide variety of real-life problems related to health care, finance, transportation, defense, weather forecasting, etc.

With the advancements in Artificial Intelligence, Humans will have a harder time understanding and retracing the algorithm's steps to a decision. The entire calculating process is transformed into a "black box" that is impossible to understand. These black-box models are built from raw data. One of the major issues with the traditional AI approach lies in the implementation of machine learning techniques. That is, one cannot blindly trust the prediction or the output of the machine learning model as that might have drastic consequences.

What is a solution to this problem? Explainable AI (XAI). It addresses the challenge of establishing trust in machine learning models. XAI stands for Explainable Artificial Intelligence. Explainable Artificial Intelligence (XAI) is a collection of methods and tactics that enable human users to understand and trust the outcomes and productivity of machine learning algorithms. The term "Explainable AI" pertains to the foreseeable impact of a model and its potential biases. In AI-assisted decision-making, it contributes to the calculation of model correctness, fairness, transparency, and results. The dimensions of an organization are crucial when it comes to incorporating AI models into operational use. An organization's adoption of a responsible AI development approach is also aided by AI's explainability.

It is an emerging artificial intelligence approach. It is also known as transparent artificial intelligence. It indicates that in XAI, one must be able to understand how and why the algorithm makes decisions or predictions. In other words, the system can justify the result that it produces. Within the realm of explainable AI, outcomes or solutions are comprehensible to humans. This is in contrast to the opaque methodology of machine learning, where even the creator or developer of the model is unable to elucidate the rationale behind specific decisions made by the AI system. Explainable artificial intelligence delivers overall data about how an artificial intelligence program decides by disclosing the merits and demerits of the program or a model, the specific criteria that has been used by the program to produce the result. It also assists in understanding why a program produces a specific result as opposed to its substitutes, which induces a level of trust that's proper for many types of decision, what type of error the program is prone to, and how the error can be modified.

There are various benefits of understanding how an AI-enabled system has arrived at a certain conclusion. Explainability can help developers ensure that the system is working as planned, it may be necessary to meet regulatory standards, or it may be crucial in allowing those who are affected by a decision to challenge or change the decision. To provide clarity on this absence of consensus, it would be resourceful to cite D. Gunning's definition of the term Explainable Artificial Intelligence (XAI): "XAI will create a suite of machine learning techniques that enables human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners."

This interpretation amalgamates two concepts that require prior discussion. Nonetheless, it overlooks supplementary factors contributing to the requirement for interpretable AI models, such as causality, transferability, informative attributes, equity, and assurance.

Let's look at the existing and future scenarios. Currently, we are using an artificial intelligence technique in which the choice or suggestion generates several questions, such as why did the model do this? Why not try something new? When does the system produce 100% accurate results? When will the system fail? When can I put my faith in this machine learning model? How can I fix the system's erroneous result? However, unlike the current artificial intelligence method, explainable artificial intelligence will include an AI explanatory model in addition to an explanatory interface is employed to assist in understanding the aspect of artificial intelligence that pertains to both the reasons behind decisions and the reasons for certain decisions not being made. The explanation module and explanation interface will also assist in understanding when the system will succeed and when one must trust this system. In this book chapter, concepts such as black-box models, transparency, XAI tools and techniques, and many more topics will be covered in detail.

2 What Are Black-Box Models?

These black-box models are shaped by a machine learning algorithm directly from data, which implies that no one, even the developers, knows how variables are joint to produce forecasts [1]. Even if one knows a list of input variables, black box predictive models might be such intricate functions of the variables that no social can understand how the variables interact to produce a final forecast. Figure 1 shows the basic diagram of black box.

Interpretable models, which share the same mathematical equivalence as black-box models but can be more ethically sound, differ in their approach by constraining themselves to offer a more profound comprehension of prediction mechanisms. In certain instances, when a compact and valid logic encompasses only a handful of variables, or when utilizing a linear model where variables are assigned weights and combined, the connection between variables and the ultimate forecast can be exceptionally lucid. Decomposable models are frequently employed to generate easily understandable models, or additional limitations are introduced to impart a heightened level of insight. In contrast, many machine learning models prioritize high predictability on static datasets over readability [3] (Fig. 2).

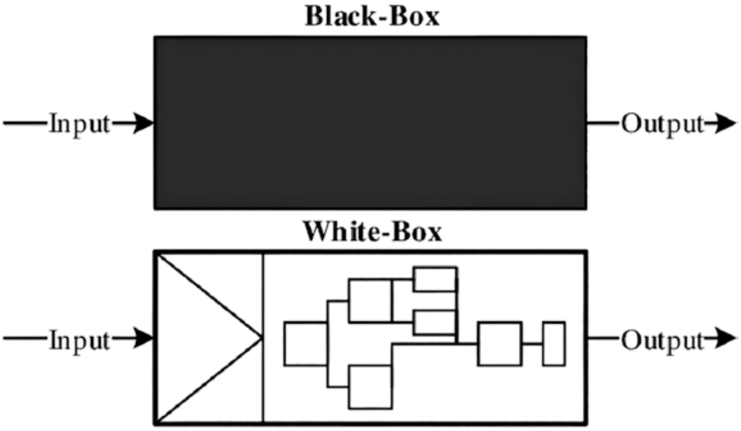


Fig. 2 Black box vs white box model [2]

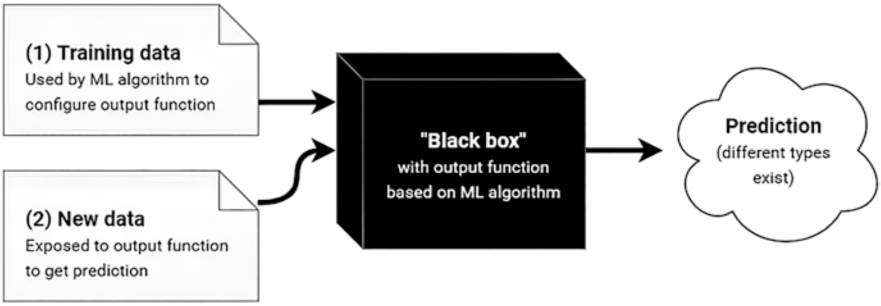

(1) Training data
Used by ML algorithm to configure output function
(2) New data
Exposed to output function to get prediction
"Black box"
with output function based on ML algorithm
Prediction
(different types exist)

Fig. 1 What are black-box models? [2]

Conversely, certain machine learning models are not designed with the intention of overcoming interpretational challenges; their primary purpose is to deliver precise forecasts based on fixed data entries that might or might not mirror the model's practical application.

2.1 Why Is Model Interpretability Important?

The interpretability of a machine learning algorithm depends on how easy it is for humans to grasp the procedures that it takes to arrive at its results. Earlier, Artificial Intelligence (AI) algorithms were known as "black boxes," with no means of knowing what was going on inside and making it impossible to explain the results to regulators and stakeholders.

It is wrong to believe that accuracy must be compromised for interpretability. When extremely basic interpretable models exist for the same tasks, it has allowed corporations to promote and sell proprietary or sophisticated black-box models for high-stakes judgments. As a result, it permits the model's developers to profit while ignoring the negative repercussions for the people who are affected. A minority of individuals contest these models, as their developers hold the view that complexity is a prerequisite for accuracy. The Explainable Machine Learning Challenge of 2018 provides an illustrative example of evaluating the advantages and disadvantages of opaque models compared to transparent models.

When utilizing the results of an algorithm to make high-stakes judgments, it is critical to understand which factors the model took into consideration and which it did not. Furthermore, if a model is not easily interpretable, the company may not be able to use its insights to make process adjustments lawfully. In tightly regulated sectors like banking, insurance, and healthcare, understanding the factors that contribute to anticipated outcomes is critical in order to comply with regulations and industry best practices.

For a variety of other reasons, interpretability is essential. For example, if researchers do not grasp how a model works, they may have trouble translating their findings to a larger knowledge base. Interpretability is also necessary for avoiding embedded bias and debugging an algorithm. It also assists scholars in decisive the impact of trade-offs in a model. More concisely, as algorithms play a larger role in society, knowing how they arrive at their conclusions will become crucial gradually.

Currently, scholars must compensate for inadequate interpretability through judgment, expertise, observation, monitoring, and careful risk management, which includes a full grasp of the datasets they utilize. Regardless of the machine learning model, there are a number of strategies for improving interpretability.

2.2 Using Explainable AI to Decipher Black-Box Machine Learning Models

Machine Learning (ML) and Artificial Intelligence (AI) have surged in popularity, finding utility across diverse sectors. However, they have also encountered escalating critique due to concerns about the reliability of their decision-making. Certain ML systems, especially Deep Neural Networks (DNNs), are often labeled as enigmatic entities because comprehending their inner workings post-training proves arduous. This opacity impedes a full grasp and explication of a model's reasoning process. Nevertheless, the provision of explanations is indispensable to establish the dependability of a model's predictions. This assumes paramount importance when machine learning algorithms underpin decision support systems in sensitive domains. Explanations not only corroborate the precision of a model's prognoses but also play a pivotal role in preempting inadvertent errors and unearthing potential biases. Furthermore, they facilitate a comprehensive comprehension of a model, which is imperative for prospective enhancements and rectification of its limitations.

Explainable AI (XAI) tackles the quandary of furnishing explanations for models that surpass human understanding due to their intricacy. These explanations span from individual (local) explications elucidating specific outcomes of black-box models—such as unraveling the rationale behind a denied loan application or an erroneous image classification—to collective (global) explanations that unveil broader patterns within such opaque models. These comprehensive explanations can address queries like identifying the most influential risk factor for a certain type of cancer.

3 Transparency in Machine Learning Models

Transparent machine learning is introduced as a new kind of machine learning which explains itself. This means that it tells us how it works, its predictions, its insights—so that the user can understand and trust the outcome. If addressed, this technology might be the best-case scenario for AI system safety and security in the future [4].

Models created by current machine learning (ML) techniques are difficult or impossible to comprehend. Security, safety, and prejudice are all issues that these deployments face. Insight into the automated decision-making process is also difficult with opaque models.

Transparent machine learning aims to tackle these issues by creating understandable models and data. It would accomplish this by displaying and altering source code representations. Consequently, you would have a possibly self-contained executable that could be used right away.

It is critical that Transformational Machine Learning (TML) systems use well-known programming languages and data formats that are simple to comprehend. Furthermore, the source code and data it generates in those languages and formats must be clear enough for an engineer of acceptable competence to understand and modify it. This is a fundamental principle that should take precedence over all other factors, even if it means sacrificing model efficiency. Later, recommendations will be made on how to achieve both efficiency and readability without permanently abandoning either.

3.1 Long-Term Objectives

Transformational Machine Learning (TML) is primarily oriented toward enhancing one or more dependently typed programming languages that possess robust specification support. This would involve the incorporation of elements from the comprehensive deep specification initiative, which strives to meticulously validate the complete developmental continuum, spanning from applications to the operating system and extending down to the hardware level [5].

What will the weather be like tomorrow?
Tomorrow will be 70 degrees and cloudy

Fig. 3 Transparency in machine learning models [6]

Furthermore, the development of source code adheres to long-term quality aspirations, including:

- Support for multiple language targets
- Comprehensive and concise commenting
- Incorporation of high-level abstractions
- Mitigation of unnecessary complexity
- Utilization of accelerated hardware for improved performance

3.2 Short-Term Objectives

The immediate objectives are:

1. Develop a transparent machine learning method that works.
2. Has it generated readable source code?

Achieving a viable proof-of-concept poses challenges. Identifying Transparent Machine Learning (TML) systems that match or surpass the performance of leading Machine Learning (ML) models would serve as a positive initial step. This rationale is supported by the potential to invigorate research enthusiasm. Nonetheless, prioritizing comprehensibility remains paramount; disregarding this aspect would undermine the project's objective, as incomprehensible source code resembles another variation of an opaque model. As depicted in Fig. 3, the illustration portrays transparency within ML models.

3.3 Theoretical Limits

It is crucial to distinguish between program readability and program comprehension ease. This does not mean that the usual criteria of reasonable competence established in the TML definition is irrelevant. Its purpose is to promote discussion of TML source model theoretical restrictions at the intersection of maximal model complexity and perfect human understanding.

Two complementary definitions of AI's foundations will be examined to aid discussion:

- *Inexplicability*: There is no explanation that is both 100% correct and understandable to humans for some judgments made by an intelligent machine.
- *Incomprehensibility*: Certain intelligent system judgments will have a 100% true explanation that no human can fully comprehend.

To address these statements, we must first explore the difference between *opaque* and *transparent* machine learning. The explanation is inextricably linked to the model in TML because the model is the explanation. Also included in that model might be a description of the TML system.

3.3.1 Layer-Wise Relevance Propagation (LRP)

Layer-wise Relevance Propagation (LRP) is a method that gives potentially complicated deep neural networks like explainability and scalability. It works by applying a set of specially developed propagation rules to propagate the prediction backward through the neural network.

3.3.2 Counterfactual Method

The counterfactual impact evaluation approach enables for determining how much of the observed real change (e.g., a rise in income) may be attributed to the intervention's influence (since such improvement might occur not only due to the intervention but also due to other factors, e.g., overall economic growth).

3.3.3 Local Interpretable Model-Agnostic Explanations (LIME)

The term LIME stands for Local Interpretable Model-Agnostic Explanations, representing key aspects of the explanation process. “Local fidelity” refers to the objective of ensuring that the explanation faithfully reflects the classifier's behavior in the vicinity of the instance under prediction.

3.3.4 Generalized Additive Model (GAM)

A Generalized Additive Model (GAM) in statistics extends the concept of a generalized linear model. Within a GAM, the linear predictor is intricately connected to smooth functions of specific predictor variables. The primary emphasis lies in making inferences regarding these smooth functions. GAMs were conceived by Trevor Hastie and Robert Tibshirani to amalgamate the advantages found in both generalized linear models and additive models.

$$g\left(E\left(Y\right)\right) = \beta_0 + f_1\left(x_1\right) + f_2\left(x_2\right) + \cdots + f_m\left(x_m\right) \tag{1}$$

In this framework, a solitary response variable (Y) is linked to particular predictor variables (x_i). The distribution governing Y belongs to the exponential family, encompassing distributions like normal, binomial, or Poisson. A link function (g), such as the identity or logarithmic function, establishes a connection between the predicted value of Y and the predictor variables. The model structure incorporates functions (f_i) that can be parametrically defined, perhaps as polynomials or unpenalized regression splines of a variable. Alternatively, these functions can be estimated non-parametrically, taking the form of smooth functions and relying on non-parametric techniques. To illustrate, within a conventional GAM, the function $f_1(x_1)$ might leverage techniques like scatterplot smoothing, such as locally weighted means. Conversely, $f_2(x_2)$ could involve a factor model associated with x_2. This adaptive nature enables non-parametric adjustments, minimizing assumptions about the genuine relationship between outcomes and predictors. While this adaptability can potentially enhance data fitting when compared to entirely parametric models, it does come at the expense of simplified interpretation.

3.3.5 Rationalization

AI rationalization involves the creation of explanations for the behavior of autonomous systems, simulating human-like reasoning. In this context, we introduce a rationalization technique that employs neural machine translation to convert an autonomous agent's internal state-action representations into everyday language. To validate our approach, we implement it within the Frogger gaming environment. Our objective is to train an autonomous game-playing agent to express its chosen actions using natural language. To build the training dataset, we utilize insights from human players who articulate their thoughts while playing.

We advocate for the adoption of rationalization as a strategy for generating explanations and present the outcomes of two studies that assess its effectiveness. The results underscore the efficacy of neural machine translation in generating rationalizations that faithfully capture agent behavior. Furthermore, the findings suggest that rationalizations are more favorably received by humans in comparison to alternative explanation methods.

3.4 Framework and Tools

Explainable AI is a new and developing discipline in the realm of artificial intelligence and machine learning. It's critical to establish human trust in AI models' choices. It's only conceivable if the dark box of machine learning models is made more transparent. Explainable AI frameworks are programmed that create reports on how a model works and attempt to explain how it works. Now we'll talk about six AI frameworks that are easy to understand.

3.4.1 SHAP

SHAPley Additive Explanations, commonly referred to as SHAP, is an abbreviation for SHapley Additive Explanations. It serves as a versatile tool for elucidating various machine learning algorithms, ranging from fundamental ones like linear regression, logistic regression, and tree-based models to more intricate models like deep learning models used in tasks such as image classification, captioning, and even in NLP tasks like sentiment analysis, translation, and text summarization. This approach is model-agnostic and harnesses Shapley values derived from game theory to illuminate model behaviors. Essentially, it unveils how diverse attributes impact the model's output and the role they play in shaping the ultimate outcome. This concept is visually represented in Fig. 4.

3.4.2 LIME

LIME is short for Local Interpretable Model-agnostic Explanations. While it shares similarities with SHAP, it boasts greater computational efficiency. LIME provides a collection of explanations that elucidate the contribution of each attribute in predicting outcomes for specific data samples, visualized in Fig. 5. Notably, LIME is versatile enough to handle any black-box classifier with two or more classes. The classifier simply needs to furnish a function capable of processing raw text or a numpy array, delivering the probabilities associated with each class. It's worth noting that Scikit-learn classifiers are already integrated with this capability.

3.4.3 ELI5

ELI5, an abbreviation for "Explain Like I'm 5," is a Python library crafted to simplify the troubleshooting and explication of machine learning classifiers. It extends its support to various machine learning frameworks, including but not limited to scikit-learn, Keras, XGBoost, LightGBM, and CatBoost.

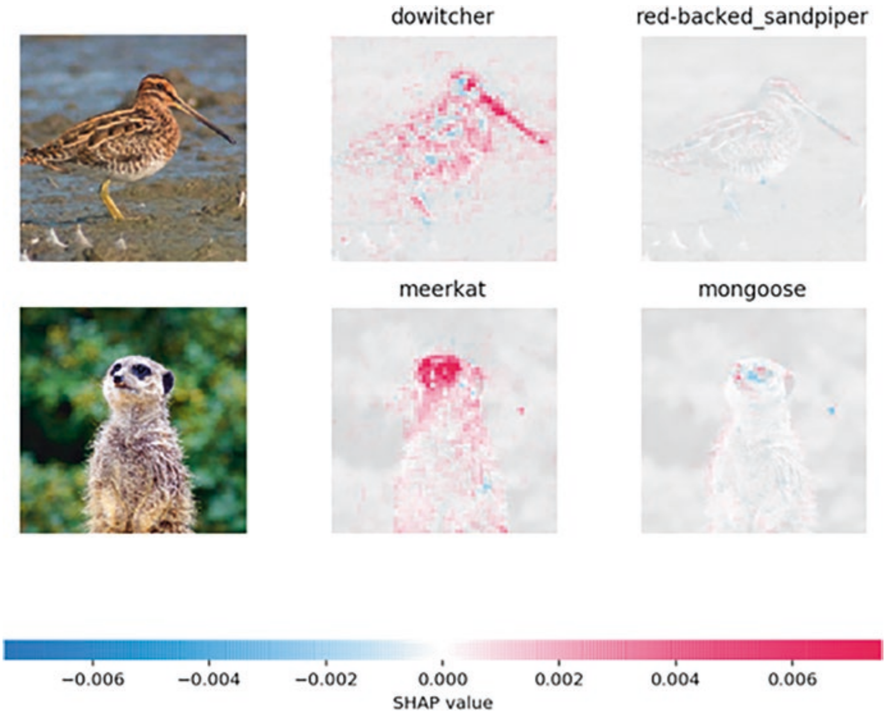
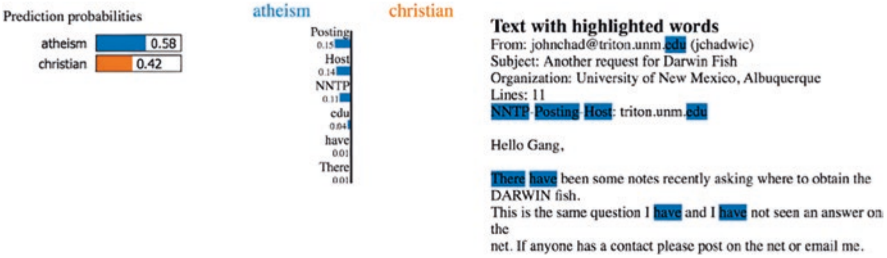


Fig. 4 Example of image classification [7]



Text with highlighted words
From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to obtain the
DARWIN fish.
This is the same question I have and I have not seen an answer on
the
net. If anyone has a contact please post on the net or email me.

Fig. 5 Screenshot of explanations given by LIME [8]

3.4.4 What-if Tool

The What-if Tool (WIT), developed by Google, serves the purpose of enhancing the understanding of how machine learning models operate. WIT empowers users to simulate scenarios, assess the significance of distinct data attributes, and visually comprehend model behavior across diverse models and subsets of input data. This tool is also proficient in handling several machine learning fairness measures. Available as an extension within Jupyter, Colaboratory, and Cloud AI Platform notebooks, WIT covers tasks ranging from binary classification and multi-class classification to regression. It gracefully accommodates various data formats, spanning tabular, image, and text data. Remarkably, it seamlessly integrates with SHAP and LIME methodologies, while also maintaining compatibility with the Tensor Board.

3.4.5 AIX360

AIX360, which stands for "AI Explainability 360," is an open-source toolkit innovated by IBM Research. Its primary objective is to facilitate the comprehension of how machine learning models make predictions. AIX360 employs a versatile array of techniques, all intended to cater to different stages of the AI application lifecycle.

3.4.6 Skater

Skater emerges as a comprehensive framework devised to aid in Model Interpretation across a spectrum of model types. Its purpose is to assist in creating interpretable machine learning (IML) systems, an essential component for real-world applications. As a freely available and open-source Python package, Skater endeavors to unveil the inner structures of black-box models on both a global scale (through extensive datasets) and a local scale (concerning individual predictions).

Here, a tabular summarization of all the frameworks implemented is shown in Table 1.

4 Evaluation Methods and Metrics for XAI

4.1 What Is the Need of Evaluation?

While dealing with artificial intelligence system we know that results are not fully accurate, and with XAI we see why the decision must have been taken that may be true or false, to solve this only we need certain methods to evaluate XAI which ultimately evaluates the XAI methods.

Saliency explanations for models have been introduced to shed light on the pertinent components of inputs that contribute to a particular decision made by a deep neural network. These explanations are part of the broader domain of Explainable AI (XAI) techniques. However, it's noteworthy that a significant portion of existing XAI methodologies still lack transparency. At present, the assessment of XAI methods draws heavily from research within interpretable machine learning (IML), concentrating on comprehension of models. This includes techniques like contrasting with established attribution methods, sensitivity analyses, reference sets of features, adherence to axioms, and visual depictions of images.

Table 1 Summarization of all the frameworks [9]

Technique	Simple to use	Stability	Efficient	Trustworthy	Feature
LIME	+	–	–	+	• Applicable to text, image, and tabular data • Employ a straightforward model for explanations, although complexity should be predetermined
SHAP	+	–	–	NA	• Built upon a theoretical framework rooted in the Shapley value concept
ELI5	+	+	+	+	• Python library that aims to provide simple and intuitive explanations for machine learning models
What-if Tool	+	+	+	+	• It provides an interactive interface for exploring model predictions and their outcomes, allowing users to gain insights into how models make decisions
AIX360	+	+	+	+	• A toolkit created by IBM, available as open-source, with a primary focus on furnishing an extensive array of tools and algorithms aimed at improving the interpretability and transparency of machine learning models
Skater	+	+	+	+	• It focuses on leveraging model abstraction and skeletonization techniques to provide insights into how models make predictions

4.2 General Steps for Evaluation of XAI

1. The training data is learned by the model. After that, the trained model predicts the test data, and the accuracy is determined.
2. A to-be-evaluated XAI approach is considered, and an explanation is generated for each sample of test data. The test data is updated by the assessment and verification procedures indicated above based on the time point relevance of the explanations.
3. The model predicts each of these freshly formed test sets, and the quality measure is calculated for comparison. Then, as a function of time, relevant change, and random changes, the quality measure/accuracy is compared. That is how we choose the outcome.

4.3 Evaluation of Methods on Time Series

A time series is an ordered collection of data points, each corresponding to a specific moment in time. Due to their size and complexity, raw time series data can be challenging to interpret, even for experts in the field. Consequently, evaluating, and validating explanations related to them presents difficulties. An assessment based solely on raw data and explanation examination is not feasible.

It is worth noting that a substantial portion of proposed Explainable AI (XAI) techniques primarily focus on black-box methods tailored for images. These methods are often fine-tuned for specific input types, such as images, and they include approaches like saliency maps. Hence, evaluating these techniques from a visual perspective holds greater relevance. In the realm of computer vision, perturbation analysis stands out as a widely-used approach. This technique involves altering a few pixels in an image (e.g., setting them to zero) based on their relevance, either the most or least relevant pixels. However, when dealing with time series tasks, assigning a value of zero can potentially indicate an anomaly. Consequently, evaluating XAI methods for time series necessitates the adoption of specialized methodologies.

Two specific methods tailored for time series data merit attention. These methods are designed to account for the sequential nature of time-oriented data.

4.3.1 Perturbation on Time Series

Here perturbation is done in two steps:

1. *Perturbation analysis*: It shows preliminary baselines for comparison. The assumption is based on the time series t and the relevance generated by the XAI approach, which is denoted by r. If r_i exceeds a particular threshold e, a time point t_i is altered. Because some time-series samples are problematic for XAI algorithms, the threshold only changes a small number of time points. The time point is set to zero or the inverse in the case of time series, resulting in two new time series (t^{zero} and t^{inverse}).
2. *Perturbation verification*: A random relevance r_{r} is employed to confirm the assertion. The number of modified time points, which is equal to the amount of r_i greater than the threshold e, is the same as in the previous example, in order to maintain the same classifier prerequisites. Like perturbation analysis, such as $t_{\mathrm{r}}^{\mathrm{zero}}$ and $t_{\mathrm{r}}^{\mathrm{inverse}}$. This technique generates new time series. The XAI approach with the random relevance method assumes that the model's quality metric q_{m} maximizes.

4.3.2 Sequence Evaluation

Two sequence-dependent strategies are used to ensure that the model and the XAI method include time series properties such as slopes and minima. Even if the perturbation analysis assumptions are correct, there is still no way to evaluate trends or patterns in the time series. The following are the evaluation methods:

1. *Swap time point*: For a time, series, and relevance, time point gets reversed after $r_{i>}$ threshold (e) the reverse time series is then added to time series.
2. *Mean time points*: Same as above just the differences take mean of sub sequence and exchange it with time series.

4.3.3 Using Canonical Equations to Get the Accuracy of Evaluation Methods

The prevalent assessment approaches have flaws, such as failing to measure accuracy in support of accountable judgments and failing to highlight where the existing methodology fails. Ground-truth data and benchmarks are commonly used in other fields. In XAI and IML, data representing ground-truth explanations is rarely employed. To address these issues, canonical equations have been developed that can be used to assess the XAI method's accuracy.

A dataset is created from an existing process, which could be either natural or artificial. Subsequently, this dataset undergoes classification to identify different classes. These classes are then represented within the dataset using mathematical equations, establishing a linkage that allows various explainable Artificial Intelligence (XAI) approaches to produce explanations. To illustrate this process, three distinct datasets are formed: “loan,” “distance,” and “time.” Afterward, a neural network model is trained using a chosen XAI metric, which could be input variance, ROAR, Post-hoc accuracy, and others. The trained model is then fed into the XAI method, which encompasses techniques such as LIME, SHAP, and others. For example, consider the case of LIME, a feature attribution method. It operates under the assumption that the behavior of a sample point can be explained through linear regression involving the point itself and its adjacent points. The resultant values are utilized for model predictions, leading to the creation of a fresh dataset. This process entails the computation of cosine similarity and the fitting of linear regression using a hyperparameter known as the number of samples. LIME is subsequently applied across the entire dataset, encompassing attributes like “loan,” “time,” distances, and the neural network model.

The data generated to represent Global Training Examples (GTEs) offers a distinctive approach to conveying explanations. However, before this data can be employed, it is essential for it to adhere to a uniform format or to have undergone consistent preprocessing procedures. This standardization ensures that the data representing GTEs can be utilized to evaluate any target XAI procedure. The resulting outcomes must then be converted to the appropriate format required by the specific technique in use. By adopting the same Ridge regression approach and regression parameters employed by LIME, particularly the hyperparameter "number of samples," linear functions are fitted. The results are subsequently measured using evaluation metrics like Euclidean distance, implementation invariance, and measure of order.

5 Challenges of XAI

5.1 Introduction

Artificial Intelligence (AI)-based algorithms, particularly those based on deep neural networks, are changing the way we approach real-world jobs performed by humans. Machine Learning (ML) algorithms have witnessed an increase in their application in automating different aspects of scientific, business, and social processes in recent years [10]. The increase is due in part to an increase in research on a branch of ML known as Deep Learning (DL), in which hundreds (if not billions) of neural parameters are trained to generalize on how to do a certain job.

The widespread use of deep learning algorithms in fields such as healthcare, ophthalmology, developmental disorders, autonomous robots and vehicles, image processing classification and detection, speech and audio processing, cyber-security, and many others demonstrates the breadth of their application. Deep learning providers can research, test, and operate ML algorithms at scale in small edge devices, smartphones, and AI-based web-services via Application Programming Interfaces (APIs) for broader exposure to any applications; thanks to access to high-performance compute nodes via cloud computing ecosystems, high-throughput AI accelerators to improve performance, and access to big-data scale datasets and storage.

5.2 Challenges to Achieve Deep Learning

While much progress is being made in the field of XAI, there are still numerous difficulties to overcome before achieving explain ability in DL models. First, there is a lack of consensus on the language and several meanings of XAI. For example, the phrases feature significance and feature relevance frequently refer to the same idea. This is especially evident in visualization methods, where there is no consistency behind concepts such as saliency maps, salient masks, heatmaps, neuron activations, attribution, and others. Because XAI is still in its early stages, the community lacks standardized nomenclature. There is a trade-off between interpretability and correctness, i.e., between the simplicity of the information provided by the system about its internal functioning and the exhaustiveness of this description. Whether the observer is an expert in the subject, a policymaker, or a user with no machine learning skills, intelligibility does not have to be at the same level in order to offer the audience with an understanding. This is one of the reasons, as previously said, creating objective measurements on what constitutes a solid explanation is a challenge in XAI. Taking inspiration from experiments in human psychology, sociology, or cognitive sciences to generate objectively persuasive explanations is one way to eliminate subjectivity. Relevant facts to consider while developing an explainable AI model include: First, explanations are more effective when constrictive, which means that a good explanation must not only show why the model made decision X, but also why it made decision X rather than decision Y. It is also mentioned that in order to create a satisfying explanation, probabilities are not as important as causal linkages. Given that black-box models prefer to process data quantitatively, the probabilistic outcomes would need to be translated into qualitative conceptions including causal relationships. Furthermore, they claim that explanations are selective, which means that focusing simply on the primary causes of a decision-making process is adequate. It was also demonstrated that the use of counterfactual explanations can assist the user in understanding a model's choice [10].

Combining connectionist and symbolic paradigms appears to be a promising approach to addressing this difficulty. On the one hand, connectionist methods are more precise, but they are also opaque. Symbolic approaches, on the other hand, are widely regarded as less efficient, but offering higher explain ability and so meeting the above-mentioned conditions:

- Because of the capacity to relate to established reasoning principles, symbolic approaches can be constructive.
- The usage of a KB that has been formalized, for example, via an ontology, might allow data to be handled directly in a qualitative manner.
- Being selected is more difficult in connectionist models than in symbolic ones.

Given that a successful explanation must have an influence on the user's mental model, which is the representation of external reality using symbols among other things, it appears that the symbolic learning paradigm is well suited to providing explanations [11]. As a result, neural-symbolic interpretability may be able to deliver convincing explanations while preserving or improving overall performance.

As previously stated, a fully explainable model should not rely on users to provide explanations since alternative explanations may emerge based on their past knowledge. A semantic representation of information can help a model provide explanations (in plain language, for example) that combine common sense reasoning with human-understandable features.

Furthermore, it appears vital to make an attempt to formally formalize evaluation procedures until an objective metric is adopted. One approach could be to draw inspiration from the social sciences, for example, by being consistent in the selection of evaluation questions and the population sample used.

A last issue that XAI approaches for DL must address is delivering explanations that are understandable to society, policymakers, and the legal system as a whole. In particular, providing explanations that involve non-technical expertise will be critical for dealing with misunderstandings as well as developing the social right to the (yet inexistent) right to explanation in the EU General Data Protection Regulation (GDPR).

6 Ethical Issues Regarding XAI

In addition to the openness and accountability of ML algorithms, ethical AI is a key topic that has drawn the attention of AI researchers who claim that ethical judgments should be one of the main drives in AI development and adoption. Researchers found multiple situations in which AI systems displayed racial bias, such as imposing harsher jail terms on black offenders or discriminating against non-white mortgage applicants. Driverless automobiles are one of the notable instances that have prompted academics to advocate for rules that govern how AI makes judgments because people's lives depend on them. Furthermore, the lack of transparency and accountability, as well as the systematic violation of people's privacy, are examples that emphasize the challenges and the need for ethical AI.

Advocates for ethical AI claim that in order for AI models to be responsible, its algorithms and models should prioritize fairness, openness, and privacy in their design. In order to achieve this aim, the FAST principles should be addressed when designing an AI project. These ideals stand for justice, accountability, sustainability, and openness. Fairness refers to algorithms as well as data, and human characteristics must be created to fulfil the discriminating non-harm standard. Accountability is focused on constructing AI systems capable of responding to dubious judgments caused by AI algorithms. The notion of sustainability ensures that AI-enabled technologies have revolutionary impacts on individuals and society. Finally, transparency provides a foundation for the AI system to explain, in simple terms, the factors that were considered while behaving in a specific way, and to justify the ethical permissibility, discriminatory non-harm, and public trustworthiness of the outcomes and the process that led to them. These principles demonstrate how AI model interpretability must be addressed while taking into consideration data privacy, model secrecy, fairness, and accountability needs. It is believed that in order for developers and organizations to create, adopt, and deploy AI approaches responsibly, these principles must be researched together.

Artificial intelligence has the potential to greatly improve workplace productivity while also supplementing the work that people can do. When Artificial Intelligence (AI) takes over routine or dangerous activities, it frees up human labor to focus on tasks that need creativity, empathy, and other abilities. When people are doing something they like, their happiness and work satisfaction may increase.

Artificial intelligence has the potential to revolutionize healthcare by boosting monitoring and diagnostic capabilities. By increasing the operations of healthcare facilities and medical organizations, AI can reduce operational costs and save money. Big data, according to McKinsey, may save the medical and pharmaceutical industries up to $100 billion every year. In patient care, the true impact will be felt. It will be life-changing to be able to construct individualized treatment plans and pharmaceutical regimens, as well as provide clinicians with increased access to data from several medical institutions to better advise patient care.

6.1 Legal Issues Regarding XAI

Is it really necessary to psychoanalyze robots? According to popular belief, we do. As artificial intelligence becomes more prevalent in our lives, many regulators and courts are beginning to insist that it be "explainable" to some extent. In layman's words, this means that the legislation requires AI users to specify [12]:

1. What information the AI has inferred about specific topics and
2. How the AI has deduced it.

The issue is surfacing in a variety of legal contexts:

People increasingly have the right under data privacy rules to know how corporations keep and process personal information about them and sometimes even the specific information that companies have on them. People can also object to automated decision-making in some cases (fancy jargon for "computers computing things"). As a result, regulators and businesses are attempting to figure out what they need to do and how to identify the information that AI has accumulated about a person.

When confronted with claims of prejudice, courts routinely ask corporations and people to justify themselves. A defendant in a discrimination litigation, particularly in employment circumstances, will need to explain what considerations prompted an employment decision and perhaps more importantly, what factors did not inspire an employment decision. When AI algorithms are used to screen resumes, set bonuses or other similar choices, concerns regarding how the AI arrived at a certain result are sure to arise.

Courts are increasingly being asked to explain how they employ AI. Many courts, for example, are increasingly utilizing AI technologies to forecast recidivism, which is subsequently included into a judge's sentence decision. The nature of judging, at least under common and civil law systems, compels judges to offer reasons for their conclusions; the responsibility is enforced even more strongly when a judge's decision deprives persons of their freedom. As a result, there is a heated discussion over how judges should explain the AI that they utilize when sentencing criminal offenders.

There are technical reasons to be skeptical of explainable AI. AI algorithms do not always mimic human cognitive processes. Really sophisticated AI algorithms, such as those based on machine learning or neural networks, may create a slew of "coefficients" to quantify this, that, or the other. However, the AI algorithm frequently defines what “this,” “that,” and “the other thing” are. A person may be unable to comprehend what these coefficients signify.

However, even if we have explainable AI operating technically, there are more difficult theoretical and legal problems that must be addressed before we go all-in on mandating AI systems to explain themselves.

First, defining AI is difficult, which makes determining which algorithms require explanation difficult. A recent Congressional attempt defines AI so broadly that it appears to include practically everything—unhelpful. On the other hand, there’s the so-called “AI Effect,” which states that when some AI technologies become widespread, people cease referring to them as AI. A typical example is optical character recognition (OCR). OCR is reliant on a slew of technologies that are, strictly speaking, based on machine learning and other AI-related technologies. Everyone exclaimed in the 1990s, “Wow, computers that can read!” Today’s technology is drab. Even when OCRed evidence is utilized in a court case, no one questions how OCR works; it is simply assumed. Likewise, almost everyone today would agree that face recognition technology is an example of AI. By the way, there is another area where people are asking that choices be explained.

Second, we’re not particularly good at explaining things. Could you do it if someone showed you a snapshot of a bird, asked you what it was, and then asked you to explain how you knew what it was? Could you name every piece of information you’ve ever heard about what a bird is? Could you describe how your brain connected these to form the notion “bird”? Could you perhaps clarify how you know it wasn’t a bat or a bug?

Third, even if we knew what variables a human brain examined and how it weighed each aspect, we don’t have a good legal framework for evaluating that judgment. In situations involving discrimination, for example, courts frequently engage in byzantine discussions over what it means for a judgment to be “based on” or “because of” something else. A criminal sentence may be vacated by an appeals court if the trial judge did not thoroughly consider a specific issue, but a different sentence may be affirmed if the trial judge studied a complex matter in a footnote. And no one I know believes that courts are truly consistent in determining whether an administrative agency has provided sufficient reasons for an adjudication or regulation. Bottom line, even when human brains make judgments, we lack dependable legal frameworks for determining when such decisions are based on the appropriate (or erroneous) criteria, and for determining how much explanation of the decision is sufficient.

Finally, even if we could overcome the technological, theoretical, and legal challenges associated with compelling AI to explain itself, will explain ability requirements do more damage than good? There are at least two unforeseen outcomes that spring to mind. People may misinterpret explain ability criteria in an attempt to reverse engineer how AI systems function. This might pose security problems. It might also lead to corporations stealing and profiting from competitors’ AI advances. On the other hand, organizations may feel compelled to simplify AI systems in order to make them more understandable, even if doing so reduces the algorithms’ effectiveness or efficiency.

7 Applications of XAI in Real-Life Sectors Such as Healthcare, Transportation, Finance, Military, Security, Legal Judgment Etc.

Technological improvements have undoubtedly provided us with better and more convenient services. Technology has become an integral component of our daily life. Its benefits far outweigh its drawbacks, and whether we like it or not, its influence on our lives will only grow. Computers, the Internet and mobile gadgets have all made our lives much easier and more efficient.

The rise in data-rich cyber-physical systems has created new opportunities to use Artificial Intelligence (AI) approaches to harness data and support the aforementioned business goals. Manufacturing, on the other hand, frequently requires certification of both the products and the procedures that produce them. This means that an AI system's automated decisions in the workplace must "explain its decision-making basis" and demonstrate that the results are traceable and reproducible. In such cases, the field of Explainable AI (XAI) is critical for the AI system's successful implementation [13].

The main objective of XAI is to develop a shared human and machine understanding of the business process, thus allowing us to comprehend what the AI system has really learned. This will allow us to better understand specific features of the AI system, such as why that forecast was made, is the prediction accurate, what are the AI system's "stable operating conditions," when is it likely to fail and so on. If we can answer these questions successfully, we'll be able to build trust and confidence in AI systems' decision-making abilities over time.

The rapid progress in the field of explainable artificial intelligence (XAI) is propelled by two significant trends. Firstly, the remarkable advancements in modern machine learning methods, notably deep learning and reinforcement learning, have raised considerable expectations for their potential in various sectors, including industry, commerce, and society. Secondly, there is a growing emphasis on establishing trustworthy AI systems, which has led to the formulation of regulatory principles aimed at ensuring the reliability and credibility of AI systems. These intertwined factors have resulted in a remarkable surge of research activity, all focused on developing a diverse array of tools and techniques to meet the demand for XAI.

However, despite this extensive research, there is still a notable absence of a comprehensive and principled framework that aligns with the historical context of explainability in scientific literature. This absence hinders the establishment of a solid foundation for creating a transparent XAI framework, highlighting the need for further advancements in this direction.

7.1 Role of Explainable AI in Different Industries

Any use case that has a direct impact on people's lives and may be affected by bias. For example, the cases discussed below are examples of the decisions made by the machine. Many AI systems are created for industrial use with features of XAI incorporated into all of them. The following are some examples of such applications:

- An AI system that not only forecasts the performance of an aircraft wing structure, but also gives the design drivers for that performance.
- An artificial intelligence system that not only anticipates an anomaly inside an aircraft wing structure, but also points to its location using a heat map overlay.
- We believe that by demonstrating more similar examples in the future, we will be able to generate greater trust and confidence in the AI system, resulting in the successful adoption of a robust AI system in the industry.
- Explainable AI plays different roles in different industries. As AI advances, we should expect to see a rise in new applications that rely less on human decision-making and accountability.

7.2 Healthcare

The potential benefits of AI in healthcare are significant, but the risk of an untrustworthy AI system is even greater. It goes without saying that AI models' recommendations to help clinicians categorize crucial diseases using structured parameters or unstructured data such as medical imaging have far-reaching implications. If an AI system predicts and also explains why it came to that conclusion, it will be far more beneficial than if it predicts and then allows clinicians to spend an equal amount of time (with or without AI judgments) determining whether the AI system's decision is accurate and trustworthy. Because lives are on the line in healthcare, XAI is critical [14].

A machine powered by explainable AI could save medical staff a significant amount of time, allowing them to focus on the interpretive work of medicine rather than a repetitive task. They could see more patients while also giving each patient more of their attention. The potential value is enormous, but it necessitates the traceable explanation that explainable AI provides. Explainable AI allows a machine to assess data and draw a conclusion while also providing a doctor or nurse with decision lineage data to understand how that conclusion was reached and, in some cases, draw a different conclusion that requires the nuance of human interpretation.

7.3 Banking, Financial Services, and Insurance

This industry has a lot to gain, and XAI has the potential to revolutionize it. Customer acquisition, agent productivity, claims prevention, underwriting, customer service, cross-selling, policy adjustment and improving risk and compliance are several of the potential use cases for AI in insurance. In banking and insurance, Artificial Intelligence (AI) has been widely used to analyze credit risk. With pay-as-you-drive and pay-how-you-drive models that leverage machine learning for decision-making, premium calculation based on multimodal information is gaining traction in industrialized countries. However, widespread awareness in developed countries about preventing data misuse has resulted in concrete regulations such as the (EU Data Protection Regulation) GDPR, which includes Article 22 on restrictions on fully automated decision-making and Articles 13–15 on the right to seek explanations for decisions made (though not explicitly stated). Because AI systems used in risk assessment, premium estimation, and other decisions are black-box models, this has a substantial influence. XAI systems that are capable of producing excellent results and delivering clear explanations will gain adequate trust and satisfy regulatory standards, resulting in increased use of AI solutions in the sector. Denying a loan, inflating/deflating premium costs for health/motor insurance or making incorrect stock trading recommendations all have larger financial stakes, therefore XAI might be seen as a useful addition in explaining such judgments to the impacted party.

7.4 How Can Explainable AI Be Implemented in Banks and Finance Industry?

Client Onboarding Process: Inadequate customer onboarding processes cost financial firms millions of dollars. Many banks find it difficult to assess their financial condition when applying for a loan. Explainable AI is a technique for determining eligibility and managing risk while keeping transparency.

Credit Decisioning: To anticipate the creditworthiness of a customer's financial institution, they are currently utilizing machine learning models that use structured or unstructured data to improve accuracy and efficiency. Because of the rise in biased AI systems, such as those based on gender or race, there is a growing demand for explainability for a fair and governable AI framework. With built-in capability for governance, auditability, and maintenance, Akira AI ensures that systems continue to run as intended.

Risk Management: AI helps banks and financial organizations track fraud and signals of potential misconduct in advance by using historical structure and unstructured data. To control risk and give a better customer experience, an AI system provides an automated risk management system.

Crime Management: Money laundering and fraud are becoming more common, putting more pressure on banks. AI enables insights to track and recognize suspicious activity in order to combat these crimes. It aids in spotting them and preventing them from carrying out their actions.

7.5 Automobiles

Autonomous driving has been a developing subject and is the industry's future. AI will undoubtedly play a part in the burgeoning field of self-driving automobiles, and explainable AI will be critical. Self-driving autos or driverless cars are exciting as long as no mistakes are committed. In this high-stakes AI application, one erroneous action can cost one or more lives.

Automakers may trace the data set by monitoring and explaining how the model got from decision point A to point Z with established data governance within the organization, making it easy for them to judge whether or not these outcomes translate to the ethical position they plan to take as a company. Passengers can also assess whether they are at ease traveling in a car that is geared to make specific decisions.

It is critical how the AI system is built and how it interacts with the vehicle. It could mean the difference between life and death. Explainability is essential for understanding the system's capabilities and limits prior to deployment. Understanding the weaknesses of driving assistance (or auto-pilot) in the field when utilized by customers is critical in order to assess, explain and rectify the issues as quickly as possible. To a significant extent, assistive parking and voice assistants are appealing features that involve the model making relatively low-risk decisions. However, in some circumstances, such as brake assistance or self-driving, XAI becomes critical in identifying and correcting bias."

7.6 Manufacturing

When it comes to detecting and correcting equipment faults in production, field technicians frequently rely on "tribal knowledge." (This has ramifications in other industries as well.) The problem with tribal knowledge is that it can vary dramatically over time: people come and go, and their information changes with them, which isn't necessarily recorded or conveyed.

Natural-language processing powered by AI can aid in the evaluation of unstructured data such as equipment manuals and maintenance standards, as well as structured data such as historical work orders, IoT sensor readings and business process data, in order to determine the best prescriptive guidance the technician should follow. This does not negate the value of tribal wisdom, nor does it eliminate the need for human decision-making.

Rather, it is a frequentative and interactive process that ensures knowledge is both stored and shared in a useful manner.

In this example, the user is presented with a number of potential repair guide recommendations based on AI, as well as the percentage confidence interval for each response as the most likely answer. On each, the user has the choice to upvote or downvote, which aids in the continual learning process and enhances future recommendations. This method allows the user to make an intelligent judgment between options rather than receiving simply one answer. The user is presented with the knowledge graph output with the input used during the AI training phase for each recommendation as an advanced feature to help the user understand the parameters on why the result was prioritized and rated suitably.

7.7 Judicial System

In western countries, AI systems are increasingly being used in the legal process for decision-making. ProPublica has exhaustively documented the inherent bias that comes with it towards a single ethnic community in the past. Bias in AI applications, such as giving parole based on the likelihood of repeat offense, has far-reaching repercussions, and fairness in them is essential because it involves an individual's rights and liberties.

AI in Criminal Justice—Machine learning and AI-powered tools can provide profound insights into humans. These technologies are particularly good at recognizing patterns that we humans often overlook. To identify solutions and create predictions, they utilize algorithms to examine massive volumes of data.

Because of AI's predictive nature, it has become increasingly popular in criminology, law and forensics, and forensic psychology. Today, algorithmic risk assessments are widely used in law enforcement. Predictive policing and assessments are being used to predict whether or not someone will appear in court or commit a crime.

Furthermore, AI techniques have been implemented in practically every level of the criminal judicial process in several countries. Bail, sentencing, and parole decisions are all influenced by algorithms. When it comes to pretrial risk assessments, many judges have resorted to predictive analytics to help them decide whether to incarcerate or release a suspect pending trial.

Some have argued that AI and predictive analytics have the ability to fix a broken criminal justice system. Many legal professionals, technologists, and community activists, on the other hand, feel that these tools may potentially exacerbate the issues they are supposed to tackle."

8 Human-Computer Interaction (HCI) and XAI

8.1 History of Interaction

Going back in time to see how interaction/communication began.

1. Make use of gestures and situations.
2. Make use of speech (most efficient way of interaction).
3. Make use of writing.

8.2 Definition

> *Interaction between humans and computers is such a way that it is seamless, secure, and simple.*

8.3 Application

In today's world, the most user-friendly systems are Windows and Android because everyone now depends on these systems. The most impactable reason behind the use of this system is a "USER-FRIENDLY" [15].

8.3.1 Why HCI Is Important

- User-centered design is becoming increasingly important.
- Increasing competitiveness via HCI research is becoming increasingly crucial today (Norman, 1990).
- Investing in high-cost e-transformation.
- Poorly designed products and services waste users' time.
- Users may even abandon using a terrible interface due to inefficient resource utilization.

8.3.2 Goals

- A fundamental objective of HCI is to improve user-computer interactions by making computers more usable and responsive to the user's demands.
- One of HCI's long-term goals is to develop systems that reduce the gap between a human's cognitive model of what they wish to do and the computer's comprehension of the user's job.

8.3.3 Theory

HCI is itself a multi-disciplinary subject that includes many concepts. It includes psychology, cognitive science, ergonomics, sociology, CS engineering, business, and graphic design. Let's discuss this field briefly for now.

We investigated how the human mind works, its functions, its behavior, its problem-solving skills, and reasoning skills in psychology and cognitive science. Understanding this allows us to create a product that is more efficient to interact with.

Ergonomics refers to the process of creating or organizing a computer or equipment such that people may interact with it in an efficient, simple, and productive manner.

Sociology is a structure or function that describes how human society is organized. Teaches us how interaction is done between human-human, human-computer.

Computer science (CS) engineering is the core part of the HCI. We have to get all the core knowledge of CS engineering.

Because we grasp all of the essential knowledge but cannot promote our products, the product will fail. We get into the business field.

Because developing the user interface for the user must be easily legible and handleable, graphic design is another important component of HCI.

Still, there are many other skills like technical writing skills, and many other fields are added in the below diagram. Figure 6 shows the disciplines in the field of HCI.

9 Current Challenges and Future Opportunities in XAI

Explainability stands as a critical challenge that holds significance both in the realm of science and within society, occupying a central role in the ongoing research endeavors within machine learning and AI. Several challenges are intrinsic to Explainable AI (XAI):

1. *Bias*: Ensuring the absence of biased perspectives acquired by an AI system from the training data, model, or objective function poses a question. Additionally, the potential influence of biases from human developers, whether conscious or unconscious, adds complexity to the matter.
2. *Fairness*: The task of guaranteeing fairness in the judgments made by an AI system is intricate. Fairness takes on diverse meanings contingent on the contextual data fed into machine learning algorithms.
3. *Transparency*: Individuals' rights to demand comprehensible explanations for AI judgments, delivered in accessible language, raise questions. The avenues for challenging these judgments and the contexts for such challenges are points of concern addressed by XAI.

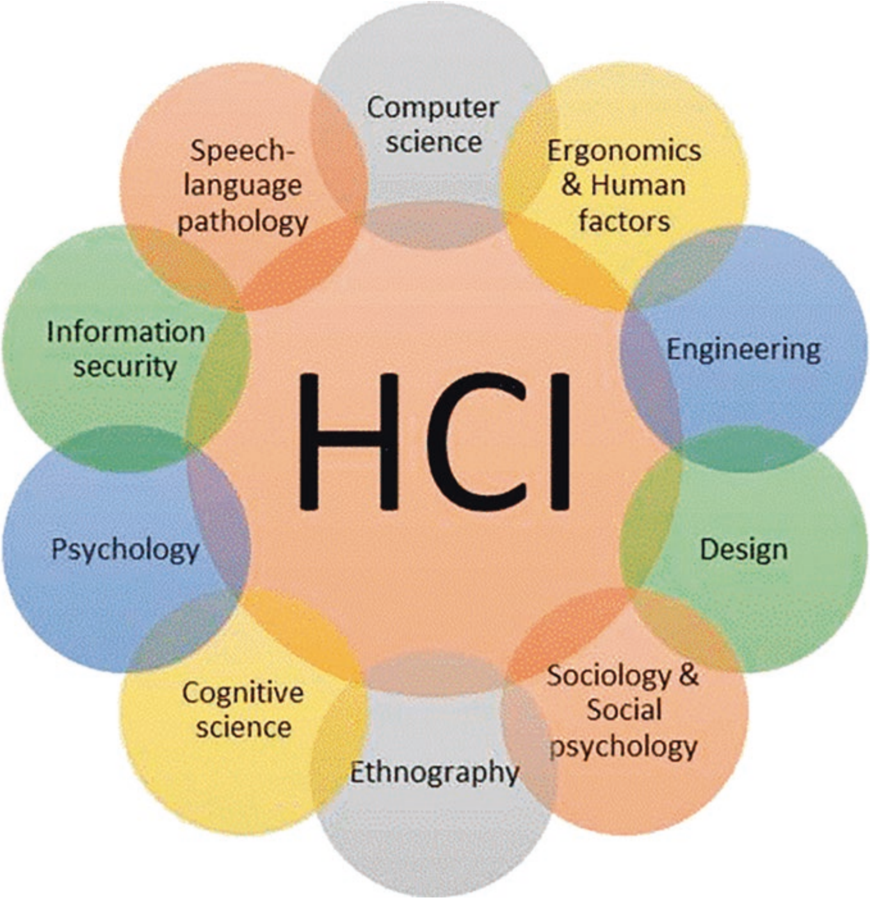


Fig. 6 Disciplines in the field of human computer interaction (HCI) [16]

4. *Safety*: Trust in the reliability of an AI system becomes challenging when it refrains from disclosing the basis for its deductions. This challenge links with the fundamental concept of generalization within arithmetic knowledge theory—specifically, the extent to which we can constrain errors associated with hidden information.
5. *Causality*: The potential for a trained model to provide not only accurate inferences but also explanations for the underlying phenomena is an essential question. The capacity for users to genuinely understand the mechanics governing a trained model remains a point of exploration.
6. *Engineering*: Effectively diagnosing and debugging inaccuracies in a trained model's outputs to enhance its performance is an ongoing engineering challenge.

In summation, these challenges underscore the intricate nature of explainability within the AI landscape. They span across technical, ethical, and practical dimensions, collectively shaping the trajectory of AI research and its pragmatic implementation across various domains.

9.1 Opportunities

9.1.1 Promoting Future Innovation Hinges on Collaborative Efforts

The recently established partnership on AI seeks to unite academics, developers, and consumers in a concerted endeavor to ensure that AI technologies yield benefits for both individuals and society. The primary objective of this alliance is to address and mitigate challenges and concerns pertaining to "the safety and trustworthiness of AI technology, as well as the fairness and openness of systems." Serving as a dynamic platform, it brings together organizations across various levels to collectively address the multifaceted challenges, complexities, and nascent issues inherent in AI. The overarching aim is to collectively advance toward mastering the intricate field of computational intelligence. This initiative goes beyond the confines of issues exclusively tied to deep learning or machine-oriented concerns.

9.1.2 The Quest for Effectively Managing Artificial Intelligence (AI) Technologies that Surpass Human Capabilities in Specific Activities Takes Center Stage

In the current landscape, we have AI systems that excel in designated domains, a trend that is projected to continue. However, this achievement sometimes comes at the cost of lacking comprehensible explanations for their operations. Consequently, it becomes crucial to undertake rigorous studies and research aimed at establishing a comparable framework for AI systems, particularly in critical deployments. This strategic approach empowers us to harness the benefits of deploying certain technologies even before achieving a complete understanding of their underlying mechanisms.

9.1.3 Elevating Decision-Making Practices by Infusing Them with a Systematic and Transparent Ethos Emerges as a Pivotal Objective

The realm of Explainable AI (XAI) stands as a transformative force that significantly elevates the quality of decision-making processes while concurrently fostering accountability among the relevant stakeholders. This leads to the formulation of meticulously crafted system requirements that engineers can effectively create, continuously monitor, and meticulously evaluate. These requirements are meticulously tailored to align with the specific domains in which they find application. As we increasingly rely on automated decision-making tools, a unique opportunity emerges: the ability to precisely define and methodically systematize the fundamental principles or values that underpin and influence our judgments.

10 Conclusion

In conclusion, the comprehensive exploration of *Explainable AI: Foundations, Methodologies, and Applications* has shed light on the pivotal role that explain ability plays in the realm of Artificial Intelligence. This research paper has meticulously examined the foundational principles that underpin the development of Explainable AI (XAI), highlighting its significance in enhancing transparency, accountability, and user trust. The various methodologies discussed within this paper, ranging from LIME and SHAP to neural machine translation and perturbation analysis, illustrate the diverse strategies available for unraveling the intricate inner workings of complex AI models.

Furthermore, the paper underscores the practical applications of XAI across different domains, including but not limited to healthcare, finance, and autonomous systems. These real-world scenarios exemplify how the integration of explain ability not only aids in comprehending model decisions but also facilitates error detection, bias mitigation, and model refinement. The insights garnered from this research paper serve as a compass guiding researchers, practitioners, and policymakers towards the responsible and ethical deployment of AI technologies. As we advance into an era where AI systems are increasingly integrated into our daily lives, the principles elucidated in this paper offer a robust foundation for fostering transparency, accountability, and user confidence. The symbiotic relationship between AI and its human users can truly flourish when AI's decision-making processes are made intelligible, ensuring a harmonious collaboration that contributes positively to society's progress.

References

1. Andrews R, Boyne GA (2010) Capacity, leadership, and organizational performance: testing the black box model of public management. Public Adm Rev 70(3):443–454
2. https://pub.towardsai.net/
3. Ljung L (2001) Black-box models from input-output measurements. In: IMTC 2001. Proceedings of the 18th IEEE instrumentation and measurement technology conference. Rediscovering measurement in the age of informatics (Cat. No. 01CH 37188), 21 May 2001, vol 1, pp 138–146. IEEE
4. Strobel M (2019) Aspects of transparency in machine learning. In: Proceedings of the 18th international conference on autonomous agents and multiagent systems, 8 May 2019, pp 2449–2451

5. Broersma M, Harbers F (2018) Exploring machine learning to study the long-term transformation of news: digital newspaper archives, journalism history, and algorithmic transparency. Digit Journal 6(9):1150–1164
6. https://developer.nvidia.com/
7. https://dev.to/article
8. https://towardsdatascience.com/
9. https://www.researchgate.net/
10. Das A, Rad P (2020) Opportunities and challenges in explainable artificial intelligence (xai): a survey. arXiv preprint arXiv:2006.11371
11. Antoniadi AM, Du Y, Guendouz Y, Wei L, Mazo C, Becker BA, Mooney C (2021) Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. Appl Sci 11(11):5088
12. Taruffo M (1998) Judicial decisions and artificial intelligence. In: Judicial applications of artificial intelligence. Springer, Dordrecht, pp 207–220
13. Longo L, Goebel R, Lecue F, Kieseberg P, Holzinger A (2020) Explainable artificial intelligence: concepts, applications, research challenges and visions. In: International cross-domain conference for machine learning and knowledge extraction, vol 25. Springer, Cham, pp 1–16
14. Adadi A, Berrada M (2020) Explainable AI for healthcare: from black box to interpretable models. In: Embedded systems and artificial intelligence. Springer, Singapore, pp 327–337
15. Xu W (2019) Toward human-centered AI: a perspective from human-computer interaction. Interactions 26(4):42–46
16. https://www.engati.com/