Innovative Research Journal of **Artificial Intelligence** (IRJAI)
DOI: https://doi.org/10.62497/irjai.239

Research Article

Why are Deep Learning Predictions Greek to me? From Explainability towards Ethical AI

Rizwana Qazi*
Public Policy and Governance from Crawford School of Public Policy, ANU, Australia; Certified Artificial Intelligence Consultant (CAIC), United States Artificial Intelligence Institute

Citation: Qazi R. Why are Deep Learning Predictions Greek to me? From Explainability towards Ethical AI. Innov. Res. J. AI. (IRJAI). 2026;4(1). Available from: https://irjpl.org/irjai/article/view/239

Article Info
Received: April 18, 2026
Revised: June 21, 2026
Accepted: June 24, 2026

Keywords
Explainable AI, Deep Learning, Deep Neural Network, Black-box, Ethical AI, Responsible AI, Trustworthy AI, AI Policy and Governance

Abstract

With a rapid proliferation of Artificial Intelligence (AI) and advanced shift from Machine learning (ML) towards Deep Learning (DL), its reliance on critical decision-making has increased exponentially. The benefits of AI are colossal particularly in education, finance and healthcare; however, the black-box nature of its deep learning algorithm has raised manifold questions regarding explainability of its decisions. This in turn, casts doubt upon authenticity and accuracy of its decision-making process. Explainable Artificial Intelligence (XAI) offers a solution to the black-box problem of the deep learning algorithm. Explainable AI has evolved with the transition from rule-based learning towards more complex neural network predictions. Enormous contribution has been rendered by academics and researchers for the development of numerous Explainable AI techniques as a post hoc explanation mechanism. However, these techniques exhibit profound limitations and impediments. It is imperative to ponder upon as why humans need explanations in the first place. An exploration of the topic further illustrates the direct relationship between core ethical values such as transparency, accountability and reliability with Explainable AI. Arguably, Explainable AI supports and upholds the necessary elements for formation of an Ethical AI system. Future directions should therefore focus on an interdisciplinary approach. The collaborative endeavor of designers, experts, developers, users, academics, researchers, ethicists, philosophers and policymakers may conjointly address the problem. Besides providing solutions of technical nature, Explainable AI may be considered as one of the enabling factors for an Ethical AI Policy and Governance framework.

** Corresponding Author:*
Rizwana Qazi
Public Policy and Governance from Crawford School of Public Policy, ANU, Australia; Certified Artificial Intelligence Consultant (CAIC), United States Artificial Intelligence Institute
Email: rizkaz786@gmail.com

Introduction

The exponential growth of Artificial Intelligence (AI) has experienced ubiquitous penetration engaging all paces of lives round the world. With the evolution of AI from machine learning model towards deep learning system discernable paradigm shift has occurred. Gradually, the role of an algorithm has increased manifold in decision-making processes particularly in the critical fields ranging from healthcare, public service, automated system, finance and education. The list of application of AI is mammoth. This exponential reliance in turn should be a sufficient proof of its reliability and trust, however this statement contains plethora of imbedded interrogations. The research article evaluates that despite increased reliance on deep learning systems, its reliability remains a daunting challenge due to the lack of transparency of a complex algorithm [1].

Remarkably, a deep learning model is a complex neural network with capability of prediction and identification of the patterns in large data sets through multiple hidden layers. The model with its multilayer's functionality contains high performance competence. However, the complexity renders model opaque in its decision explainability due to its black-box nature [2]. Even the decision-making process is often not visible to its developers or designers [3]. This lack of explainability of deep learning model has raised several concerns regarding fairness, trust and accuracy. Trust is a critical element for relying on model's predictions for taking any decisions or deploying models and without understanding the reason of such decision, the model may not be trusted for deployment, as actions have consequences [4][5]. This is particularly significant when entities rely on machine learning outcome for reaching crucial decision-making processes. This article explores that due to the complex nature of deep learning architecture, lack of the interpretability of decision-making logic poses a problem. The opaqueness of the model may render those decisions prone to bias, discrimination and unjust treatment raising significant risks on individual's right to access to job, services, opportunities or fair medical diagnosis or treatment. Consequently, the opaque nature of deep learning may give rise breeding grounds for an ethical breach including lack of transparency and accountability [3].

The problem of black-box has been addressed through various studies and technical solutions [4]. Numerous explainability techniques have been devised to improve the transparency of models. Together these techniques build the framework of Explainable AI. Despite the fact that significant progress in resolving black-box has been on-going, yet most literature views Explainable AI as a technical challenge. Though black-box problem is often considered as a technical one, yet it presents a broader challenge for accountability and transparency in decision-making processes by the AI system [5]. The real dilemma arises because the decisions made by the deep learning system impacting individuals cannot be understood, scrutinized, contested, questioned or justified resulting in untoward repercussions. The central question remains why is Explainable AI viewed as a technical problem requiring technical solution; can it be understood in relation to its consequences which are similar when it comes to an ethical consideration?

The major contribution of this research article therefore, lies in raising concern about the daunting problem which is inherent in a complex algorithm. Since most of the scholarly literature on Explainable AI is technical in nature, hence main scope and objective of this research article is to provide simple account for the policy makers to understand the risk of lack of transparency, accountability and unfairness in the AI system due to black-box nature of algorithm. The article is not based on any scientific foundations, rather it presents narrative description from policy perspective and possibility of inculcating Explainable AI in an ethical framework. Though, explainability techniques provide technical solutions, yet development of Explainable AI can also provide an abundant opportunity for policy makers towards building an ethical AI framework. Numerous instances suggest how lack of explainability posed ethical issues. This leads us to position it at the intersection of interdisciplinary approach. Addressing deep learning explainability issues will therefore play a decisive role in ensuring that the AI system demonstrates fair, transparent and trustworthy mechanism.

The research article is structured in various sections. Section 2 is founded on discussion on why deep learning model is considered a black-box through its contextual and strategic make-up. Section 3 ponders on philosophical discussion on necessity of explanations for humans in relation to AI inculcating the existent regulations of European Union (EU) as a case for highlighting its significance. Section 4 enumerates few case examples to demonstrate how interpretability and ethical AI are connected connotations. Section 5 attempts to explain the core concept of Explainable AI, whereas Section 6 incorporates some basic differences in various approaches. Section 7 attempts to provide a simple account of few Explainable techniques from the layman perspective. Section 8 explores some limitation and gaps demonstrated by techniques along with discussion in section 9 which may be considered as impediments in implementation of Explainable AI framework particularly its challenge to imbue ethical principles. Last section 10 recommends future directions in two arenas, one through theoretical changes, whereas another suggests the possibility of an application of Explainable AI as of the essential components of an ethical AI mechanism with further research in the area.

Why is a deep learning model known as a black-box?

Briefly to recall, deep learning model is one of the sub-components of machine learning, which in turn, is the part of the broader artificial intelligence. If artificial intelligence occurs when computer mimics the human intelligence, machine learning happens when computer is predicting results based on past pattern. On the other hand, deep learning functions on the pattern of the human brain neurons. Whereas, machine learning makes predictions based on historical data and visible patterns, deep learning is founded on a complex design system. Deep learning mimics biological or human brain's neural network. It is therefore known as Deep Neural Network (DNN) [2]. The process of machine learning may be supervised or unsupervised. When data is labelled by humans and predictability is almost known, we call it a supervised learning. Conversely, unsupervised learning is not labelled, as the system is trained on large unseen data for discovering predictions based on this exposure. Unlike traditional machine model, deep learning can function on unsupervised, unlabeled and unstructured data with proficient accuracy. DNN based models can automatically handle engineering feature as there is no need for explicit feature extraction under the supervision from humans [5] [6].

As stated in the preceding paragraphs, deep learning model has taken an utmost inspiration from the human brain, which is a complex neural network and functions on similar intricate pattern. The core foundational block of deep learning is neural network. These building blocks are artificial neurons, simulating billions of neurons of the human brains. Developed in 1950s by Frank Rosenblatt as Artificial Neural Network (ANN), the neurons are known as perceptrons. The artificial neuron or perceptron contains simple information called weight and bias. They act like memories of algorithms. The architecture of deep learning is designed in a way to possess three or more layers, broadly classified as input, hidden and output layers as shown in Figure 1. The input layer passes the input data to subsequent layers. The middle layers are known as the hidden layers which contain biases and weights. The output layer functions in relation to previous layers and produces outcome. This basic neural network pattern is termed as Multilayers Perceptron (MLP). This complex system of layers works through a loss function and an optimizer algorithm. The loss function determines the error or difference between predicted and accurate answer through updating weights and biases. The bigger the loss function, the lesser the accuracy of the outcome. The optimizer function algorithm works to minimize this loss, to predict outcome which progressively comes closer to accuracy. These multilayers of deep learning model along with optimizer and other components serve as learning pattern from massive, unstructured data to predict outcome [5].

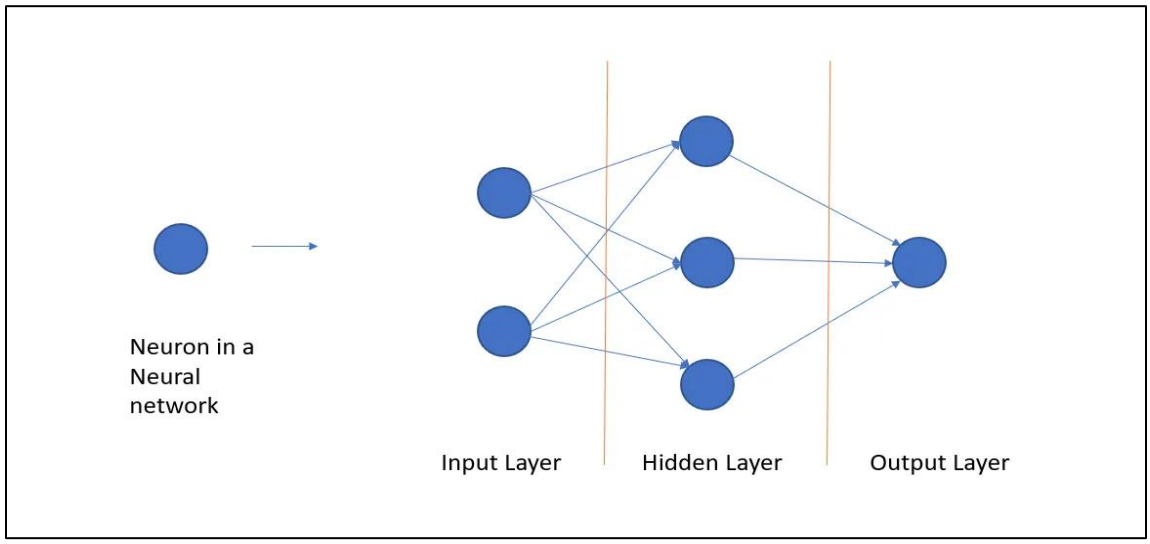

Figure 1: Deep Learning Basic Concept of Layers [7].

Due to its vast ability to automatically learn complex patterns from massive dataset, deep learning is very useful in resolving complicated approximations. The application of deep learning spans around numerous domains, including Generative AI, Computer Vision, Natural Language Processing and Image Generation [2]. Image recognition is important in healthcare, whereas finance, recruitment and public sector realms are being largely benefitted from various functions of deep learning models. ChatGPT is a popular illustration of Large Language Model (LLM). Other examples include an exhibition of an extraordinary performance of deep learning model in medical image analysis, where it has successfully been applied to various tasks such as tumor detection, disease diagnosis, radiological image classification and pneumonia identification. Previous studies demonstrate an extensive adoption of deep learning techniques in medical imaging and their capabilities to accomplish high diagnostic with accuracy, making them valuable tools for supporting clinical decision-making. Nevertheless, the complexity of deep learning model often limits its interpretability to the user [8].

As we can evaluate, though the outcome of deep neural network is incredible, yet there are rising concerns about lack of interpretability of the prediction. This lack of understanding of the prediction's outcome create a dilemma particularly in critical decision-making processes as enumerated in above paragraph. Molnar in his book Interpretable Machine Learning defines a black-box model as a system which 'does not reveal its internal mechanisms' [9]. Further elaboration reflects that black-box model 'cannot be understood by looking at their parameters' e.g., weight and bias. Deep learning model is enigmatic, when it comes to comprehend why certain outcome is predicted. It is unfathomable. Due to this reason, it persists as a notable concern [10].

Why is explanation necessary?

This study addresses that explanations are significant for the human beings for numerous reasons. Explanations build trust. The humans have tendency to trust and rely on those decisions for which they generate ample understanding in their thought-processes. For instance, a medical practitioner is more likely to trust a diagnostic system of an AI, if it explains which medical image, symptom or patient characteristics has caused some particular recommendation. Explanations can influence actions. As an example, if a patient is detected with brain tumor by AI system and it is justified explanation, then surgeon can opt for surgery or any other treatment suitable for it. The strategy to act, depends upon a valid reason which comes through solid comprehension of the matter. Explanations also help in navigating problems to find out solutions in an informed way. Explanations help in choosing and steering the action in a rational direction. They act to aid like an evidence-based

informed policy decision.

In a traditional human decision-making environment, individuals tend to lean towards a system, which provides freedom of speech and information. As opposed to the authoritarian system, the democratic architecture is built on the edifice of clear carving and identification of rights and their implementation in an organized mechanism. We live in a free and liberal society, where our rights hold as much importance as our basic necessities. Most of us would simply be eager to know the reasoning behind every decision made which somehow impacts us. The right to information is a mounting doctrine of our society particularly in this age of information revolution, which has engaged and assimilated every one of us under one united global digital platform.

The concept of individual's right to explanation has been pioneered by regulatory landscape within the European Union. The European Union's General Data Protection Regulation (GDPR) bestows upon individuals the right to get meaningful information about the reasoning of decisions which impact them. This notion is further supported by recent EU Artificial Intelligence Act. This law has strengthened the necessity of AI models to be interpretable and understandable for their explanations through introduction of accountability, transparency and human oversight in automated decision-making processes. The right to an explanation may appear to enshrine the broader connotation of individuals right to Explainable AI. This regulatory evolution demonstrates the significance of explanation particularly in high stake decisions affecting individuals, regulators and developers. These developments are not only the manifestations of growing importance of building trustworthy and Explainable AI but a legally compliant AI system [11].

Significantly relevant is the concept of agency. The AI system works as an agent because it is designed to act on behalf of a user such as humanoid robot and driverless car etc. An agency refers to the capacity of a system to perceive its environment, process information, make decision and act accordingly to achieve certain outcome on behalf of the user. For example, driverless car through its sensors perceives its surrounding and decides its action on the basis of processed information. We often use term intelligent agent for the AI system [12], as it selects actions based on learned pattern or internal reasoning without human intervention. This concept has integral significance in the area of an ethical AI, as the AI system is theoretically responsible for action without inflicting any harm to the human being [13]. This is important that the AI system is trained to ensure trust and sanctity of human values while making crucial decisions. The foundation of agency and its related decision-making capacity are based on accountability and responsibility, which cannot be attained if a model behaves like a black-box. The explanation provided by the model is the cornerstone of agent-user relationship [14].

Beyond this, humans naturally seek explanations to make sense of the world around them. The human mind functions on reasons and relies on cause-and-effect relationships to understand and experience the moments around. The interrogation about necessity of explanation also leads us to search meaning of our survival, which lies in human cognitive abilities. Yuval Noah Harari in his book "Sapiens: A Brief History of Humankind", narrates how homo-sapiens surpassed other ancient species. The homo-sapiens are progressed with cognitive abilities which make them exceptional from other species. Harari argues that on a large-scale, human cooperation is evolved by imagined realities and shared myths. The need for explanations is deeply rooted in human cognition. As argued by Harari, from ages, humans have relied on meaningful interpretations and shared narratives to understand the world and coordinate activities [15]. Consequently, when AI systems gradually influence crucial decisions, individuals naturally seek explanations for how and why these decisions are made. Humans tend to assign and find meanings in the world surrounding them, and this is the reason humans cannot settle anything less if elucidation by machine learning system is behind the black-box.

Previous studies also emphasize upon necessity of explanations [9]. Academics argue that the practical applications of deep learning are widespread [2]. Regulated industries such as medical and financial are increasingly relying on deep learning system in their affairs. Among other applications of deep learning, criminal justice and legal field occupy a

prominent position. However, black-box models present challenges for lawyers and researchers. The risks and challenges posed by the "black-box" nature of AI model in legal field is highly prone to biases and discrimination. This opaque nature of AI model remains impediment to the transparent and trustworthy decisions. Academics therefore advocate for Explainable AI with human oversight to produce reliable AI system in legal application [16]. Another study cautions that deep learning system due to its inherent complexity lacks interpretability which limits its adoption in critical domains such as healthcare, governance, e-commerce, banking and public safety [2].

The above discussion is aimed to put forth an argument that these perceptions validate that the necessity for explanation is not exclusively a technological prerequisite but also a manifestation of a cognitive, legal, and social expectation. Institutions require explanations to justify decisions and regulators rely on explanations to ensure transparency and accountability. Whereas humans seek explanations not only to establish trust but to align with logical reasons. Hence, explainability should be viewed as an interdisciplinary approach engaging various stakeholders rather than an entirely computational subject.

How is lack of explainability an ethical consideration?

On the basis of preceding discussion, this article argues that both explainability and transparency are associated. Theoretically, transparency is an antonym of opaqueness. With transparency comes the clarity and trust. These norms play key roles in accountability and responsibility matters in a democratic system. This article argues that Explainable and Ethical AI can both be viewed from the same perspective, as the goals of both are to be trustworthy and responsible towards their actions. As Ethical AI is an emerging field, and there is not universally accepted definition yet, however an Ethical AI framework, besides others, also encompasses the elementary principles of accountability, fairness, transparency and accuracy [17]. One cannot expect transparency, fairness and accuracy in the system unless its logic and reasoning of predictions are known. These principles are intertwined and keenly enshrined in foundation of any ethical system, which are also relevant objectives in the Explainable AI framework [3].This suggests that explainability functions can act as enabling condition for fairness and accountability rather than as an explainable principle in isolation.

It can be inferred from existing literature that when the deep learning model is exposed to a large data set, it is prone to attract bias, discrimination and unfairness. The prediction, as a result, may be the reflection of hidden bias and discrimination, which gives arise concern to its fairness, accountability and transparency. In high-risk decision-making environment such as healthcare, it is therefore imperative to have Explainable AI [5] [4]. The necessity of an Explainable AI has also emanated from some real-world examples. There are instances where inability of understanding algorithm decisions caused undesirable consequences. These case studies expose a shared pattern that a major ethical concern is not simply that AI systems produced biased predictions, but that the reasoning behind those outcomes remained largely incomprehensible. As a result, effected individuals could neither contest nor implicitly understand the decisions.

Back in 2012, the United States Court adopted an AI based system known as COMPAS [Correctional Offender Management Profiling for Alternative Sanctions] which was aimed to predict the likelihood of repetition of offensive actions by the criminals. The system's internal structure was complex and opaque resulting in incomprehension of the risk score generation by the judges and legal experts. Later in 2016, it was understood that COMPAS had hidden racial bias and black defendants were classified as more likely to repeat the crime. As the model reasoning was opaque, therefore, it was not possible to challenge the ruling based on AI's tools predictions. This case raised serious concern regarding discrimination and hidden bias in AI systems. This black-box nature of the model posed severe consequences on the life of the defendant, as he was unable to understand and challenge the reason behind risk scores, transparency and due process of law [18].

A notable example is an incident in 2016 by an

automated Tesla which was on autopilot. The self-driving car was unable to differentiate between brightly lit sky and white side of 18-wheeler truck. This resulted in an attempt by Tesla to plow through the truck. The passengers suffered considerable injuries. This unfortunate incident gave rise to the significance of comprehensible self-driving system for the users. The findings supported the importance of the explanation of deep learning system to avoid the untoward consequences impacting human lives [2].

An AI healthcare risk prediction algorithm was adopted by Optum a subsidiary of United Health Group in the United States to predict health outcome based on medical conditions of a patient. It was later found that an algorithm had hidden bias against black patients. The AI tool underestimated the healthcare needs of black patients as compared to white patients with similar medical conditions. The complexity and opaque nature of the algorithm made it difficult for users to detect bias in an earlier stage [19].

One prevalent case from the Dutch Childcare benefits scandal depicts the ethical impact of black-box model AI tool. Families were tagged as potential fraudsters, yet neither policy makers nor the affected individuals could understand as why was a particular person classified as high-risk or how was the risk score calculated? It was also unknown as which factors contributed to the prediction of labelling families as frauds. The underlying algorithm responsible for making its prediction did not reveal factors for assigning fraud risk scores. As a result, citizens were unjustly harmed. This discriminatory practice continued without detecting the problem at a former stage, which if priorly known could have preempted harm to the citizens [20].

What is Explainable AI?

The case studies from the preceding section reaffirms the need for Explainable AI, which has now emerged as one of the significant fields of research. Explainable AI can be referred as the set of techniques which are devised to make predictions or output of AI understandable to the humans [1]. It is usually abbreviated as XAI; however, this paper uses Explainable AI for the sake of better understanding the arguments. Some academics assign different meanings to interpretability and explainability, whereas some researchers use these terms interchangeably [9] [12]. Previous studies agree on the definition of interpretability as "the degree to which the cause of a decision can be understood to the humans". In other words, interpretability is the extent to which humans can predict the model's outcome consistently [12] [9]. There is also a direct relationship between interpretability of machine learning and human understanding. The higher the interpretability of the model, the better understanding for the humans and vice versa. Hence it can be inferred that the higher the level of understanding the better the transparency and trust in AI decisions [13].

The concept of explainability in AI has evolved through decades. Its origin can be traced back to the early expert rule-based system in 1970s and 1980s. The early symbolic AI and rule-based expert systems relied on reasoning which were based on rules and were naturally fathomable to the human understanding. For example, an expert system on medical field was built in 1970s named as MYCIN. The goal was to detect infection and suggest treatment during the diagnostic procedure. The rule-based expert system was inherently interpretable, because it could identify laboratory finding, symptom and diagnostic suggestion by relying on group of 'if-then rules' by tracing the sequence of logical deduction. However, due to manually created rules, the system exhibited uncertainty, incomplete information and lack of scalability which resulted lags in the system. The limitation in the expert system paved the way for the machine learning approach in 1980s and 1990s, which was mainly based on the data-driven solution [1].

The journey of machine learning from expert system to deep learning model with improvement in predictive performance emanated at the cost of transparency. The paradigm shift was visible with the addition of varieties of models such as linear, logistic regression, decision trees, support vector machines [SVMs] and random forests which further transitioned towards complex deep neural networks

[8]. In 2010s, the black-box nature of neural networks set the stage for Explainable AI as a separate research field. Since then, frequent explanation techniques; a few will be discussed in ensuing section; have been established to expand transparency, accountability, trust and human oversight [1].

Since then, a monumental amount of work has been performed by scholars and experts in this field. One of the sophisticated contributions in Explainable AI is combining social construct to the machine learning. Academics believe that the design of Explainable AI must be in relation to cognitive understanding of the minds, rather than totally engaging on technical front. Explanation is a social process engaging communication between machine and human. The study also contends that humans generally like contrastive explanation, while most of Explainable AI techniques are focused on internal working of the machine learning system, however it is significant that it should provide meaningful and useful explanations to humans [12]. Another significant contribution in the field of Explainable AI evaluates a systematic framework for machine learning model for explanation and interpretation. The interpretation is relevant while dealing with black-box models particularly for deep learning and ensemble models. The insights serve both as a theoretical foundation and a practical guide for Explainable AI [9].

Classification in relation to Explainable AI

There are a range of ways to classify Explainable AI. This sorting is based on certain criteria such as nature, scope, area and application of Explainable AI. Few are very briefly discussed below:

Inherent and post hoc explanation

As the name implies, the inherent explanation is considerably intrinsic to the nature of the model. As a simple analogy, something we possess by birth like a genome on DNA, or anything to have by virtue of its inherent nature. Rule-based AI models such as decision trees and liner regression are intrinsically designed in a way that internal logic is discernable to the user. This inherited explanation approach for machine learning model is referred as intrinsic or ante hoc explanation. This approach offers explanation in a transparent way for human understanding. As a drawback, inherent interpretability comes at the cost of accuracy of prediction; however, the significance of this explanation method is crucial for machine learning explainability area. Nevertheless, this approach does not posit notable challenge in explainability of AI [2].

On the other hand, post hoc explanation emanates after model yields its prediction. Due to complex internal makeup, the system behaves discreetly and does not reveal its internal logic or it depicts 'black-box' features, as earlier discussed. Deep neural network and ensemble model fall under this category. In order to extract the explanation, various techniques are employed [1]. Few of these techniques are discussed later in the article.

Local and Global Explanation

This classification is based on scope of interpretability in case of post hoc explanation. Local explanation offers clarification of an individual prediction or single feature responsible for the decision. For example, an income criterion responsible for approving the bank funds. Contrary to specific individual prediction, a global explanation covers the overall behavior of the model across an entire input space. It describes the overall logic of the model. This allows insights into model structure and feature importance [2].

Agnostic and Specific Model Explanation

Model-specific and Model-agnostic approaches are widely used terms in Explainable AI techniques. This classification is based on model application. As the name suggests, model-specific method is applicable for specific type of model. Antithetical to it, model-agnostic way is applied universally to all models irrespective of any particular model dependency [9].

Existent Explainability Techniques

This research article highlights some of the widespread explainability techniques used for explanation of deep learning predictions. The techniques enumerated in this article are depicted as representative illustrations of key explainability approaches rather than an exhaustive list of all explainability methods. Explanation techniques are broadly represented in four different approaches, such as feature attribution method, saliency or

visualization technique, counterfactual and rule-based explanation.

Foremost is the LIME or Local Interpretable Agnostic Model of Explainability. LIME supports local interpretation or specific model prediction with locally interpretable model. LIME is known as one of the earliest post hoc explanation techniques. It is one of the feature attribution methods. The key terms are Local and Agnostic which highlight that LIME is both for local interpretation and it is not specific to one model but can function in any model. The local explainability of LIME suggests that it can explain the main element responsible for making a particular decision or in other words, LIME interprets the individual prediction of complex deep learning, so it can be comprehensible the humans. For example, if a loan is refused to a person, then LIME may identify the income criteria necessary to earn the loan [if annual income is less than 24,000 USD then no loan; if income is greater than 24,000 USD, a person may be eligible]. Another main feature of LIME is its model agnostic nature, which shows that it can be applied to a variety of machine learning models such as deep learning model, support vector machine and ensemble model. Noteworthy to mention a recent advancement in LIME framework is the Sub-modular Pick LIME (SP-LIME) version for addressing the issue of local interpretability. SP-LIME provides global explanation to selected representative instances and tends to lower the gap between local and global explanation [21] [22].

Another popular feature attribution technique is SHAP which is the abbreviation of SHapley Additive exPlanations (SHAP). It is also a kind of post hoc explanation. SHAP is a very useful explanation technique as it demonstrates how much each feature has contributed to a particular decision. SHAP is based on a cooperative game theory with the concept of Shapley value which was originally proposed by Lloyd Shapley in 1950s. In game theory, Shapley value estimates how much each player contributed in the overall result of the cooperative game. SHAP provides a unified framework for interpreting machine learning predictions. For example, imagine a model is predicting house prices, then SHAP values will determine how much each feature such as house size, location, number of rooms and age of property contributed in estimating market price of the house. The fair contribution of each feature by considering all possible combinations of features is determined by SHAP. One of the major features of SHAP which makes in strong candidate for industry and research is its strength in providing both local and global explanation. SHAP also satisfies three properties including local accuracy, missingness and consistency. It works on mathematical equation which provides SHAP values quantifying how much each feature contributed towards prediction. SHAP is considered as widely used method in Explainable AI for justifying AI-assisted decisions [21].

In addition to LIME and SHAP, visualization technique such as Gradient-weighted Class Activation Mapping (Grad-CAM) has also been proposed to improve interpretability of deep learning for image building. Grad-CAM is of the post hoc interpretability techniques which uses saliency/visualization methods. Grad-CAM generates class-discriminative localization maps by utilizing gradients for identification of image regions significant for prediction. Grad-CAM is a gradient-based visual explanation technique, which though specifically designed for Convolutional Neural Networks (CNN) is widely applicable to a range of CNN based models. Relatively recent modifications are also devised as the extensions to Grad-CAM such as Grad-CAM++, Score-CAM, X Grad-CAM and Ablation-CAM. Grad-CAM is considered as one of the foundational techniques in Explainable AI due to its significant contribution in improving interpretability in deep learning model in computer vision and for important tasks such as image identification and visual classification. The main aim of Grad-CAM is to help humans understand prediction by model's visually highlighting focused regions in the image which has influenced that decision. For example, if the image is predicted by model as dog, then Grad-CAM will identify focus image region responsible for this identification. As discussed, Grad-CAM is model-agnostic so within CNN architecture, it can function in various ways. For example, in image caption model, this technique can show which image area influenced particular caption. Similarly in image classification, it can highlight particular region such as object bodies or animal faces, whereas in visual question answering, Grad-CAM can identify which image part the model

focused when answering a question [23].

Counterfactual Explanation is considered one of the noteworthy categories of Explainable AI method alongside feature attribution and saliency-based approaches. The counterfactual explanation elucidates that rather than indicating to the users why a model made a decision, a counterfactual explanation illustrates them what slight variations would have produced a diverse outcome. Rather than recognizing which features influenced a prediction, it defines the minimal changes to an input that would alter the model's output, thereby providing users with actionable and human-understandable explanation. Such explanations are particularly significant in high-stakes domains, including, finance, healthcare and criminal justice, where understanding how a decision could change is often as important as understanding why it occurred. For example, Mr. A's loan application was rejected. If his annual income had been $5,000 higher and existing debt $2,000 lower, the application would have been approved. This type of explanation is intuitive because it is actionable. In contrast to feature-attribution methods which recognize variable that influenced a prediction, counterfactual explanation defines the minimal changes required for an input to produce a different outcome [24].

One of the local explanation variant techniques is Anchor, which is a rule-based approach. It is based on conditions or shortcut rule, behaving like 'if-this-condition-is-true, then the model will give this particular answer'. Anchor functions on set of conditions or rules which if fulfilled, the model prediction is definite and predictable. For instance, an Anchor might work like 'if a person's income is above USD 100,000 and has not history of bank default, the model must always go for approval of the loan. An Anchor behaves as 'lock-in rule' fixing the model's predictions in certain situations. In other words, by generating explanations in the form of if then rules or conditions, the model's explanation becomes predictable. The user can explicitly understand why a model produced a certain output based on the meeting of specified conditions. One of the advantages of using this technique is that it does not only give high precision but has clear coverage boundaries. Anchor provides model-agnostic explanation; hence it can be applicable to any classifier or model with precision [25].

Relatively recent contribution lies in an innovative Explainable AI system by combining local explanation to the global model with a novice approach GLocalX. GLocalX is an abbreviation of Global Local Explainable AI. The key idea is to reconstruct a meaningful global understanding of the model by combining many local explanations. GLocalX model suggests that instead of developing a global explanation from the very scratch, it is better to rely on local explanations as those offer more accurate predictions around one specific data points. However, local explanations do not provide big picture of the model, so we can merge local explanations to make a global interpretable model. GLocalX is based on the philosophy of local to global. GLocalX assumes that bunch of similar instances may have similar explanations, hence local explanations may be composable into broader general explanation. It is a model agnostic explainability framework which hierarchically aggregates rule based local explanation to the global interpretability model by addressing the issue of local to global explanation [26]. Table 1 provides a brief comparison of techniques discussed below.

Table 1: Comparative Analysis of Explainable AI Techniques

Explainable AI Techniques	Strength	Limitations	Typical Application Domain
LIME	Builds a local surrogate model around one prediction, model-agnostic	May exhibit instability	Healthcare, Credit Scoring
SHAP	Measures feature contributions using Shapley values from	Computationally expensive	Healthcare, Finance

	game theory		
Grad-CAM	Uses CNN gradients and visual explanations	May exhibit limitations due to CNN	Medical imaging, object detection
Counterfactual Explanation	Demonstrates small changes required to obtain different predictions	Multiple plausible counterfactuals	Loan approval, insurance
Anchors	Uses rule-based conditions	Computationally intensive	Tabular Data

Limitations and Gaps

Despite extensive contribution has been made in the formulation of explainability techniques, yet limitations still persist. Collectively, technical approaches as discussed in above section namely LIME, SHAP, Grad-CAM, Counterfactual Explanation, Anchor, and GLocalX establish a milestone in creating black-box models interpretable. This is appreciated that Explainable AI techniques improve model transparency through various technical solutions, however these techniques also pose certain limitations. The observations which are analyzed from the discussion of explainability approaches are outlined below:

Explainability techniques largely explain model prediction after decisions have been made. Post hoc approximations explain the output and not the actual internal reasoning or logic. This observation suggests that these techniques do not transform opaque models into inherently transparent systems. Techniques which serve the purpose of providing post hoc explanations may not always reflect true model reasoning. Post hoc explanations therefore are naturally prone to errors and inaccuracy. They usually hide structural bias, which may lead to ethical concerns. Hence, post hoc explanations do not necessarily improve transparency and may truly conflict with the requirements of accountability and fairness [18].

Each technique resolves only one area of explainability and not a single technique provides a comprehensive solution to address explainability. Both LIME and SHAP largely recognize influential features, Grad-CAM highlights visually significant regions, counterfactual explanations demonstrate how outcomes could change, while anchors provide rule-based estimates of model behavior. As a result, each technique is dissimilar from others and offers partial perspective of the model [25].

One of the gaps lies in the discussion of correlation and causation, which holds considerable importance. Most of the techniques identify statistical relationship rather than causal association. Feature attribution approaches such as LIME and SHAP identify which variables contributed most to a prediction but cannot estimate whether those variables actually caused the decision. Likewise, saliency/ visualization method such as Grad-CAM indicates image regions that caused a prediction without signifying why those regions are causally related. This gap becomes particularly problematic in high-risk areas, where decisions entail causal reasoning rather than statistical correlation [27].

Most of the techniques have underlying assumption that all users are contented with one identical explanation. The implicit assumption that 'all users think alike' is prone to misleading responses from the perspectives of different users. Different stakeholders have varying ways of viewing explanations. As an illustration, computer engineers may prefer feature attribution or SHAP values, whereas policy practitioners, bankers, lawyers, judges, policymakers and clinical experts may require explanations from their own perspective and context. This leads us to the notion of usefulness. Altogether existent techniques offer explanations which deal with technical solution. However, the explanation needs to be useful to the user, which depends upon the suitability of their audiences.

The main purpose of explainability techniques is to provide explanation. However, mere presence of explanation is not the guarantee of trust. Arguably, explanation builds trust. Supposedly, if explanation is wrong, inaccurate or unstable it can miserably backfire. This scenario may lead to further loss of trust in the AI system [22]. In this situation, the system may not only be labelled as a black-box or opaque but rather an inaccurate and unreliable. As an illustration, even in normal human-to-human relation, a deceiver may be condemned more than an ignorant person. For argument purpose, the relationship of machine and human may also follow the same analogy.

Academics demonstrate phenomenon of Clever Hans, which is relevant here. For example, in vision and medical imaging, the model while relying on spurious correlations may highlight regions that appear clinically meaningful. Due to this phenomenon Explainable AI techniques may create false confidence in reasoning of the model, which can lead to over-trust in high stakes domains [28]. Consequently, model may achieve correct predictions for the wrong reasons, and explanation fails to reveal the true decision logic [18].

Grad-CAM has several strengths but limitations too. The explanation often can be ambiguous. This method does not fully explain the internal reasoning process of the model, although it shows the focused areas. Another limitation is coarseness of the heatmaps as they depend on convolutional layers with lower spatial resolution. Grad-CAM does not explain causal reasoning but only explain correlation. The assumptions of feature independence may not always hold true. Often correlated features and approximation methods may be complicated and unstable [23].

Prior literature demonstrates that anchor improves human ability to simulate model behavior compared to other explanation methods. However, anchors may not cover all instances, meaning some predictions remain unexplained if no high-precision rule is found. Due to this reason, the system can experience lags, which was one of the limitations of the traditional rule-based system.

The preservation of fidelity is another limitation. The literal meaning of fidelity is to be faithful or loyal. In machine learning, it refers to the accuracy of the model in relation to its predictions. In simpler terms, with relation to the machine learning, it refers to the explanation being faithful or technologically sound to the decision-making process. Often the explanation may be superficially looking reasonable just to placate the user, however it might not be technically correct. The faithfulness is at stake [26]. This is to create the impression of righteousness shrouded in the mystery of falsehood. We often call it as illusion of truth. In the digital sphere, it is easier to fall into this trap by trusting the false explanation.

One of the technical challenges is scalability. Machine learning systems have tendency to grow complex, which make explanations pretty hard to understand. As we understand, deep learning models have heaps of parameters and highly complex internal system of multidimensional feature space. It is often computationally infeasible and monetarily expensive when scalability of deployment comes under consideration [29].

Challenges

Despite progress in the development of explainability techniques challenges persist [5]. Explainable techniques only attempt to explain model behavior, and each method has limitations as discussed in earlier section. The methods do not automatically guarantee trustworthy or meaningful explanations and the problem remains largely unresolved. Together with limitations, it is essential to acknowledge challenges and impediments in a widespread adoption of an Explainable AI system.

Moving forward, theoretical challenges may be deliberated upon. Prevalent literature argues that despite numerous research work being evolved in the area of Explainable AI, this field lacks a shared definition of interpretability. Not only universal definition but, unified evaluation mechanism of explanation has not been finalized [2]. The exertion of a search for a universally accepted definition

poses enormous hurdles. It might happen that deep neural network is interpretable to the medical radiologist with visual explanation but not for layperson. Similarly, the linear regression model may be interpretable for the statistician but not for the policy analysts [5]. Hence, the lack of universal conceptual foundation leads to varying understandings of explanation with different system designs. The explanation may hold true for machine learning engineer but might carry incomprehensible elements to the physician or policy maker.

The universal definition of explanation also surrounds with related discussion about its characteristics. Academics in their paper published in 2017 titled "Towards a Rigorous Science of Interpretable Machine Learning" divulged upon the similar conundrum. The argument put forth in their paper points out that the definition of interpretability is not a single phenomenon, but rather it revolves around three integral concepts such as context, task and decision. The definition of interpretability is subjective and can be viewed differently from diverse perspectives. What counts as a good explanation can be observed from the user's perspective. The user can hold various positions such as a data scientist, regulator, policy maker, expert or end-user. On a similar notion, explanations can have varying meanings depending on the task being performed such as trust or fairness auditing, debugging or scientific discovery [30]. Nonetheless, the definition of interpretability should be relative to the purpose of the model. Interpretability may be understood in relevant to the humans' relations with machine. Designing of context-specific explanation system for Explainable AI offers a great challenge for academics, researchers and policy makers.

A Significant challenge is also encountered in the area of evaluation of explanations. There is no universally accepted framework to measure or evaluate the quality of an explanation. The current evaluation metrics of machine learning such as precision, recall, accuracy or FI-Score are mostly quantifiable and objective. Explanations on the other hand, have more quirks of cognitive understanding, perception, trust and usefulness. These are the traits which are not quantifiable. The evaluation standards such as fidelity, completeness, fairness, user satisfaction, comprehensibility and usefulness might seem tempting solutions to the formulation of evaluation mechanism, but inculcating all elements in machine learning systems may seem daunted. Additionally, there are chances that these dimensions may contradict each other. For instance, comprehensibility may reduce fairness or accuracy and vice versa, which can further exacerbate the challenging situation [30].

The human-centered explainability is also similar theoretical challenge. Explanations are subjective as discussed in above paragraphs. Different users may require different explanation based on what they are looking for their own perspectives, context and expertise. A doctor diagnosing a disease may want a simple computer vision explanation, highlighting the troubled areas responsible for the medical issue. A patient looking at his report might be seeking an answer in a simple natural language rhetoric to explain the symptoms of the disease along with some suggestions for its treatment. A policy maker may be interested in digging up the explanation which has fairness and ethical justifications for its speaking points for presentation in policy implementation. The lawyer may be more interested in legal explanation useful for his case. The conundrum of the diversity of the audience poses a challenge for creating an explanation which is human-centered. For one type of audience, the explanation may serve the purpose, but for others it may be deceptive [30].

We all know that real world scenarios are dynamic particularly in an environment where technology and autonomous system function. Same is true for AI system, which is highly dynamic and fast evolving area. In this backdrop, explanations generated at one moment may not hold valid, once parameters are updated. This is one of the key challenges, when consistency of explanation mechanism might fluctuate in constantly adaptive environment. There is a likelihood that explanations may not be reproducible, which may reduce their suitability for real-life contexts. Studies show that widely used methods such as LIME and SHAP can produce inconsistent explanations for similar inputs or change significantly under minor perturbations. This undermines their reliability as trustworthy tools for accountability particularly in sensitive applications of the finance and healthcare [31].

There is no denying the fact that the machines learn by correlation and its prediction based on a past pattern. On the other hand, humans seek to find reasoning in causality. Practically, machines can identify co-occurrence between variables. It can work on statistical irregularities in data set; however, it does not disclose whether one variable has actually caused another. For instance, if a model is trained on bank data for loan approval, then past history, income level and credit score might be relevant features for predicting the outcome. However, the model does not sense causality, that how these features can cause default in loan payment. Most of Explainable techniques, discussed in preceding sections are used to describe why a model made a specific decision by identifying particular features. The cause-and-effect factor has not yet been inculcated and producing the process of causal understanding and assumption of relationship based on causality may be a daunting task [5].

The trade-off between accuracy and performance is considered as a key challenge in Explainable AI. Many academics highlight the trade-off between predictive performance of the model and its interpretability function [18]. Arguably, it is often considered that the higher the model complexity is, the more accurate predictions are. It is also argued that black-box models usually attain higher accuracy because they can vastly extract abstract and nonlinear representations from large datasets. This indeed creates an anomalous scenario. Increasing model complexity may improve predictive performance but decreases interpretability for the humans. This trade-off between accuracy and explainability impact various application domains. For instance, in the domain of healthcare and law, explainability holds a great score as its predictions might affect human lives. At the same time, accuracy holds equal merits. As a result, trade-off between accuracy and interpretability poses a legit impediment [2].

As discussed earlier, one of the vital challenges arise in the area of ethical considerations of machine learning. AI system is often trained on historical data, with the likelihood of biased datasets. The probability that the model may either amplify or reproduce discrimination or unfairness, cannot be ruled out [2] [16]. The problem gets messier when biases remain undetected since the model may hide its reasoning process. Even though Explainable AI techniques discussed above such as SHAP and LIME are designed to improve trust and transparency however these methods may present new ethical risks rather than fully resolving opacity problems. The existing explanation techniques are not sufficient to uncover discriminatory patterns. This leads to an ethical conundrum of underlying biases not only in outcomes but also in explanations. This is one of the key challenges.

Future Research Path

The article bifurcates future research directions into two roads. One path is aimed for overcoming rigorous challenges concerning universal approach for consolidating fragmented themes of Explainable AI. This includes mechanism of universal thematic solution. Second path is trodden, but it is essential to evolve concern regarding usefulness of Explainable AI for the formation of an ethical AI.

The first path leads to identify universal solutions towards thematic challenges including universal definition of Explainable AI, Standard Evaluation Matrix, meaningful and human-centric explanations. The scalability of explainability techniques is one of the components.

Primarily, one of the crucial areas of research should be the development of a universal definition of Explainable AI and Standard Evaluation Matrix [2]. It is established that enormous development has been rendered for building numerous Explainable AI techniques, which are very useful. However, the body of work is fragmented. Various techniques differ in terms of accuracy, applicability and scope. The universal accepted definition and methodology for evaluating faithfulness, usefulness, quality and effectiveness of explanation has largely remained a herculean task. Future research directions should therefore focus on establishing universally recognized definition, standardized benchmarks and evaluation framework which could foster meaningful explanation to the users. Needless to say, the consistency and standardization of rules of explanations applicable across various domains is the pressing need. For adoption of Explainable AI, a

unified definition and evaluation metrics would facilitate the process across various applications.

Future research should prioritize the design of explanation techniques which are meaningful to the wide range of users. These should serve the purpose for the variety of users. Explanations intended for AI developers may vary considerably from those required by regulators, domain experts, healthcare practitioners and policymakers. Future design should focus on more adaptive and user-centered explanation frameworks. These explanation methods should not only be correct technically but must assign accurate meanings to the predictions. Human-centered explanation would be useful in informed decision making and building trust in algorithm's predictions.

One of the challenges discussed in above sections was scalability for the deployment of explainable AI in mainstream fields. Future studies should focus on building scalable explanation methods with the capacity of interpreting gradually complex machine learning models, while keeping a balance between performance and interpretability. Furthermore, research should prioritize building explainability techniques for increasingly sophisticated AI architecture such as large language and foundation models. Existing explanation methods often strive to offer comprehensible interpretations of highly complex systems. It would therefore be essential to address these limitations for ensuring accurate and scalable deployment in high-risk domains.

The second path explores policy formulation to cater explainability issues in line with an Ethical AI. The article argues that explainability should not be considered as a standalone phenomenon but rather one of the foundation elements for an ethical, responsible and trustworthy AI. Future studies should also examine the incorporation of explainability with other ethical AI principles such as accountability, fairness and transparency. It is established that Explainable AI is fast emerging field and the role of Explainable AI is touching beyond the boundaries of mere provision of explanation towards the domain of an Ethical AI governance. There is a thin line of demarcation. The implementation of Explainable AI poses considerable challenges. However, once these impediments are addressed, it can serve as a double sword by laying the foundation for the development of an Ethical AI as depicted in Figure 2, which provides a conceptual relationship between Explainable AI and key foundational principles of ethics.

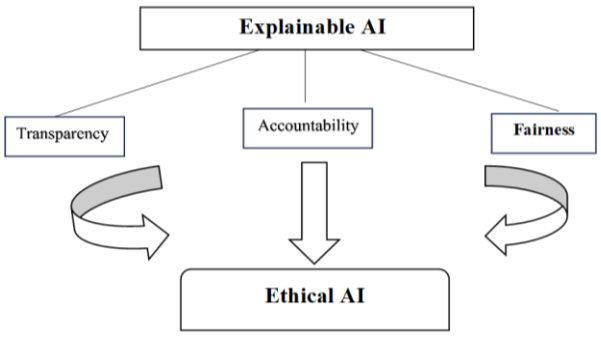

Figure 2: Conceptual relationship between Explainable AI through key ethical principles [Qazi R. Why are Deep Learning predictions Greek to me? Towards Explainable Artificial Intelligence (XAI). IRJAI. 2026;4(1):15. Available from: https://irjpl.org/irjai/article/view/239].

The future of Explainable AI should be viewed from a broader perspective encompassing not only an element of interpretability of deep learning decisions, but should inculcate the larger area of an ethical, responsible and trustworthy AI system [3]. The practical implications of Explainable AI techniques are enormous for policy makers, regulators and industry practitioner. Policy makers should consider comprehensive governance frameworks in which Explainable AI functions as a complementary system for bias mitigation and responsible AI deployment. Such integration is significantly crucial in high-risk domains [18], where algorithmic decisions have key societal repercussions. Explainable AI may be taken as a starting point to carry out the ambition of establishing an Ethical AI. The research article trusts that formation of Explainable AI should be taken as an opportunity to inculcate basic ethical norms and values in AI system.

Moreover, Explainable AI is likely to play an important role in AI Governance and regulatory

compliance. Evolving legal frameworks, including the European Union AI Act, emphasizes transparency, human oversight, accountability along with risk management as indispensable requirements for trustworthy AI systems. Future research should therefore explore how Explainable AI can internalize ethical values by garnishing public trust in AI technologies. The research article understands that though Explainable AI is necessary but it is not sufficient for an Ethical AI framework. An ethical AI framework would envisage broader connotation, but Explainable AI could be one of the pillars [13]. The future directions should ponder over questions as how researchers can build a foundation by strengthening the relationship between explainability and AI governance.

The above discussion seeks to conclude that role of AI with an advent of complex deep learning system has become more complex and significant. The transition from rule-based expert system to the neural network machine learning paved the way for Explainable AI. The role of Explainable AI, to make algorithm decisions comprehensible to the human understanding, occupies an important position in the growing age. However, challenges remain. The examination of challenges poses foundational queries for policy makers. It is indispensable to understand how AI policy and governance may find answers to the challenges and complexities involved in propositioning Explainable AI.

The future work should explore interdisciplinary collaboration among computer scientists, ethicists, researchers, scholars, philosophers and policymakers to warrant that explainability both supports and complements ethical AI governance. Such integration will benefit translating technical developments into practical governance framework to flourish transparency, accountability and public trust in the AI system. Such interdisciplinary collaboration merging technical innovation with ethical, legal, and governance perspectives, will be vital for safeguarding trust of humans in the AI system. In future, algorithm should not only be accurate but also transparent, accountable, and socially responsible. Ultimately, the future of Explainable AI lies not merely in developing algorithms more explainable but making it ethically aligned, legally accountable, and socially trustworthy. Explainable AI should not be regarded merely as a technical solution to black-box models but it warrants more collaborative research. It may provide practical implications in future in constituting an integral and enabling mechanism for achieving transparent, accountable and trustworthy AI governance which are embodiment for the foundation of an Ethical AI.

References

1. Kabir S, Hossain MS, Andersson K. A Review of Explainable Artificial Intelligence from the Perspectives of Challenges and Opportunities. Algorithms. 2025; 18: 556. https://doi.org/10.3390/a18090556
2. Hassija V, Chamola V, Mahapatra A, Singal A, Goel D, Huang K, et al. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. Cognitive Computation. 2023; 16: 44-75. https://doi.org/10.1007/s12559-023-10179-8
3. Qamar T, Bawany NZ. Understanding the black-box: towards interpretable and reliable deep learning models. PeerJ Computer Science. 2023. http://dx.doi.org/10.7717/peerj-cs.1629
4. Lundberg M, Lee SI. A Unified Approach to Interpreting Model. In 31st Conference on Neural Information Processing Systems [NIPS]; 2017; Long Beach, CA, USA.
5. Ribeiro T, Singh , Guestrin. Why Should I Trust You?". 2016. http://dx.doi.org/10.1145/2939672.2939778
6. Tal AS, Sheidin J. XAI4RE- Using Explainable AI for Responsible And Ethical AI. New York City, NY, USA: In Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization [UMAP Adjunct'25]; 2025.
7. Guidotti , Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys [CSUR]. 2019; 51[5]. https://doi.org/10.1145/3236009
8. Arrietaa , Rodr´ıguezb D, Sera D, Bennetotb , Tabikg S, Barbadoh , et al. Explainable Artificial

Intelligence [XAI]: Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. Information Fusion. 2019.

9. Qazi. Ethical Deliberations in Artificial Intelligence. Innovative Research Journal of Artificial Intelligence [IRJAI]. 2025; 3[2]. DOI: https://doi.org/10.62497/irjai.179
10. Philomena. Introduction to Deep Learning with TensorFlow and Keras. [Online].; 2023. Available from: https://python.plainenglish.io/introduction-to-deep-learning-with-tensorflow-and-keras-d9008758ba61.
11. Ahmed , Naz S, Khan S, Rehman AU, Ismael WM, Khan MA. Explainable artificial intelligence [XAI] in medical imaging: a systematic review of techniques, applications, and challengesExplainable artificial intelligence [XAI]. BMC Medical Imaging. 2026; 26[37]. https://doi.org/10.1186/s12880-025-02118-w
12. Molnar. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable: Lean Publishing ; 2019. Available from: https://leanpub.com/interpretable-machine-learning
13. European Parliament and Council of the European Union. Regulation [EU] 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence [Artificial Intelligence Act]. Official Journal of the European Union; 2024. Available at https://artificialintelligenceact.eu/article/86/
14. Miller T. Explanation in artificial intelligence: insights from the social sciences. Artificial Intelligence. 2019;267:1-38. doi:10.1016/j.artint.2018.07.007
15. Baker S, Xiang W. Explainable AI is Responsible AI: How Explainability Creates Trustworthy and Socially Responsible Artificial Intelligence. arXiv:2312.01555v1. 2023. arXiv:2312.01555v1. 2023.
16. Johnson DG, Verdicchio M. AI, agency and responsibility: the VW fraud case and beyond. AI & Society. 2019; 34: 639-647. https://doi.org/10.1007/s00146-017-0781-9
17. Harari N. Sapiens: A Brief History of Humankind New York : Harper; 2015.
18. Yu R, Ali GS. What's Inside the Black Box? AI Challenges for Lawyers and Researchers. Legal Information Management. 2019; 19: 2-13. https://doi.org/10.1017/S1472669619000021
19. Jobin A, Lenca M, Vayena E. The global landscape of AI ethics guidelines. Nature Machine Inteligence Perspective. 2019;[1]: 389-399. https://doi.org/10.1038/s42256-019-0088-2
20. Rudin. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. Nat Mach Intell. 2019; 1[5]: 206–215. doi:10.1038/s42256-019-0048-x
21. Obermeyer Z, Powers B, Vogeli C, Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019; 366[6464]: 447-453. https://doi.org/10.1126/science.aax2342
22. Hadwick D, Lan S. Lessons to Be Learned from the Dutch Childcare Allowance Scandal: A Comparative Review of Algorithmic Governance by Tax Administrations in the Netherlands, France and Germany. World Tax Journal. 2021; 13[4]. https://doi.org/10.59403/27410pa
23. Selvaraju , Cogswell , Das , Vedantam , Parikh , Batra. Grad-CAM:Visual Explanations from Deep Networks via Gradient-based Localization. Computer Vision Foundation. .
24. Wachter S, Mittelstadt , Russell C. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. Harvard Journal of Law & Technology. 2018 Spring; 31[2].
25. Ribeiro , Singh , Guestrin. Anchors: High-Precision Model-Agnostic Explanations. In The Thirty-Second AAAI Conference on Artificial Intelligence [AAAI-18]; 2018: Association for the Advancement of Artificial.
26. Setzu M, Guidotti , Monreale , Turini , Pedreschi , Giannotti. GLocalX -From Local to Global Explanations of Black Box AI Models. Artificial Intelligence. 2021; 294[103457]. https://doi.org/10.1016/j.artint.2021.103457
27. Jain , Wallace. Attention is not Explanation. In Proceedings of NAACL-HLT 2019; 2019;

Minneapolis, Minnesota: Association for Computational Linguistics. p. pages 3543–3556. https://doi.org/10.1038/s41467-019-08987-4

28. Lapuschkin S, Waldchen S, Binder , Montavon G, Samek W, Muller KR. Unmasking Clever Hans predictors and assessing what machines really learn. Nature Communications. 2019; 10: 1096.
29. Saeed W, Omlin C. Explainable AI [XAI]: A systematic meta-survey of current challenges and future opportunities. Science Direct. 2023.
30. Doshi-Velez , Kim B. Towards A Rigorous Science of Interpretable Machine Learning. 2017.
31. Melis DA, Jaakkola TS. Towards Robust Interpretability with Self-Explaining Neural Networks. Computer Science: Machine Learning. 2018. https://doi.org/10.48550/arXiv.1806.07538