FRAMEWORK FOR EXPLAINING BLACK-BOX MODELS USING EXPLAINABLE AI (XAI)

AWODELE S. O
DEPARTMENT OF COMPUTER SCIENCE,
BABCOCK UNIVERSITY ILSHAN-REMO,
OGUN STATE, NIGERIA
awodeles@babcock.edu.ng

FAYEMI T. A
DEPARTMENT OF COMPUTER SCIENCE,
BABCOCK UNIVERSITY ILSHAN-REMO,
OGUN STATE, NIGERIA
fayemi0197@pg.babcock.edu.ng

OJUAWO O. O
DEPARTMENT OF COMPUTER SCIENCE,
BABCOCK UNIVERSITY ILSHAN-REMO,
OGUN STATE, NIGERIA
ojuawo0687@pg.babcock.edu.ng

OLORUNYOMI O. B
DEPARTMENT OF COMPUTER SCIENCE,
BABCOCK UNIVERSITY ILSHAN-REMO,
OGUN STATE, NIGERIA
olorunyomi0052@pg.babcock.edu.ng

MUSTAPHA M. M
DEPARTMENT OF COMPUTER SCIENCE,
BABCOCK UNIVERSITY ILSHAN-REMO,
OGUN STATE, NIGERIA
mustapha0219@pg.babcock.edu.ng

FARUNA J. O
DEPARTMENT OF COMPUTER SCIENCE,
BABCOCK UNIVERSITY ILSHAN-REMO,
OGUN STATE, NIGERIA
faruna0100@pg.babcock.edu.ng

&

CHUKWULOBE I
DEPARTMENT OF COMPUTER SCIENCE,
BABCOCK UNIVERSITY ILSHAN-REMO,
OGUN STATE, NIGERIA
chukwulobe0408@pg.babcock.edu.ng

Abstract

Further development of Artificial Intelligence (AI) and, especially, in such complex systems as Deep Learning (DL) and Large Language Models (LLM) resulted in their widespread application to the essential fields of human activity, such as healthcare, finance, and education. Non-linear, complex designs of these high-performance models, however, make them black boxes that are hard to understand, and their inner workings and decision-making are hard to interpret. This has brought up a high concern on the issue of trust, accountability, and ethical governance. This paper evaluates how the Explainable Artificial Intelligence (XAI) can alleviate this issue by rendering black-box models more transparent, understandable and interpretable to end-users and stakeholders. XAI solves the problem of interpreting complex algorithms and making them human friendly by offering ways of describing the processing of input data and decision formation. The significance of XAI is supported by the necessity to comply with the regulation, such as the General Data Protection Regulation (GDPR) that requires accountability of the automated decision-making, as well as the requirement to be fair, to detect and reduce biases hidden in the models. The three primary approaches of XAI methods are; Model-Agnostic Post-Hoc Interpreters (MAPHI): Techniques that are used after a model is trained, such as LIME and SHAP, that explain a prediction locally or globally; Intrinsically Interpretable Models (IIMs): Models that are inherently interpretable, such as decision trees though they can be less predictive power than LLMs; Overarching Frameworks and Auditing (OFA): Governance frameworks such as Responsible AI (RAI) that enact principles like Fairness, The problems of XAI, including the inherent trade-off between model accuracy and interpretability and the threat of explanation hacking are also addressed. To solve these problems, models such as OpenXAI are being studied to standardize the technical critique of the methods of explanation in terms of such important measures as faithfulness, stability, and fairness. Finally, XAI is not merely a technical requirement but an ethical foundation of successful AI implementation, as it is necessary to make the systems more human-centred and transparent, to allow building more trust and to enable the responsible AI development.

Keywords: *Explainable Artificial Intelligence (XAI), Black-Box Models, Large Language Models (LLMs), Deep Learning (DL), Regulatory Compliance*

Introduction

Artificial Intelligence (AI) is theorization and creation of computer systems that can execute the work that conventional human intelligence, including visual perception, speech recognition, decision-making, language translation, etc., can carry out [1]. An artificial intelligence (XAI) is an element of AI, which is a learning models created to read, comprehend and interpret circumstances and make a meaningful conclusion that are more transparent, interpretable, and understandable to end users. As a number of researchers argue, XAI increases the transparency of decision-making to human [2]. Certainly, with the complexity of AI systems, especially with the emergence of Deep Learning (DL) and Large Language Models (LLM) systems, their internal processes tend to be black boxed, becoming non-transparent systems [3]. XAI helps in this problem by offering ways of explaining the process of processing input data and coming up with decisions, so that the stakeholders can comprehend and have trust in AI systems [4]. XAI is also significant as it helps to eliminate the divide between complex algorithms and human comprehensibility [5]. Against this backdrop, XAI was created to have a more interpretable and friendly automated decision-maker techniques that is easier to use by the end users in critical areas of life, such as healthcare, finance, education and law; the non-transparency of an AI system may cause mistrust and ethical issues [6]. Such as, in a case where an AI system makes a medical diagnosis (healthcare) or refuses a loan application (finance) or even criticizes equity standards in a questionnaire (education), the stakeholders must know the rationale behind the decision to be fair, accurate and responsible [7]. XAI is also crucial in the detection and mitigation of AI model biases; once the cause and effect of decision-making can be established, it will be simpler to observe the tendencies of unfair treatment or misjudgments due to bias training data. Such transparency assists organizations in following ethical standards and legal provisions which includes General Data Protection Regulation (GDPR) that requires accountability, usability and reliability of automated decision-making processes [8].

Statement of the Problem

The development of AI has brought us to the growth of knowledge and its usage in any aspect of life and science as it was identified in medicine, science, law, education etc. Human confusion muddled to identify the ideas and motivations of the decisions of the AI devices applied in various fields and fields with the concept of black-box

models because the applicability of AI tools and its result are hard to determine. The machine learning systems are becoming less and less human run, more powerful, larger, and less understandable as they are being used to make decisions with a huge real-world effect. These AI application systems must be trustworthy [2]; a layman must have the capability of comprehending the model itself, or at minimum obtain why the model has constructed a specific choice. Such situations resulted in the emergence of black-box in artificial intelligence tools and its applications to other fields. The possibility to interpret, comprehend and explain the motives of the choices of AI devices has allowed more research on XAI to justify the ethical conduct of the AI machines in responsiveness and applicability. This paper thus, evaluates the black-box models that are transparent using the Explainable AI (XAI).

Related Works

The drivers for Explainable Artificial Intelligence (XAI)

Most models developed using contemporary Artificial Intelligence (AI) and Machine Learning (ML) algorithms, and especially Deep Learning (DL) [21] are termed black-box models, which are by their nature highly complex in nature. They are viewed as a black box since it is hard to comprehend the inner logic and intricate inner working of the box [20]. The interpretability or explainability weakness in these AI systems is assumed to be a significant factor that has brought forth major concerns in the minds of researchers, practitioners, and policymakers. Black-box character is acutely felt in high-performance systems such as Deep Learning (DL) models and Large Language Models (LLMs); since the latter, e.g. in their ongoing rapid development, are black-box in nature, this raises serious challenges in terms of accountability, trust and interpretability [19]. The problem with this is that such systems present a significant delay when used in sensitive application areas, including the health sector, financial sector, education, government, and security where a single wrong move might have lifelong effects [23]. The transparency of these black-box models is lacking since its complex structure makes it hard and, in many cases, inefficient to explain the mechanism of learning and decision-making processes and, simply put, the sheer complexity of these systems has rendered their inner workings opaque such that experts can no longer easily trace the path of how the input data went to form the final derived decision [18]. This challenge in deciphering the behavior of the model restricts confidence in the reliability of the model [18], this non-transparency is a big bottleneck to further implementation of the models in high stakes areas where it is critical to understand the rationale behind a decision to ensure validation and compliance. The increasing demand of Explainable Artificial Intelligence (XAI) is based on the demanding positions of complex society, ethical, and practical demands placed on contemporary black-box AI models. These needs create a need in transparency and interpretability in many dimensions such as regulatory compliance, confidence in the user, model maintenance and knowledge advancement.

The XAI push is largely motivated by the development of new strict regulatory systems, including the General Data Protection Regulation (GDPR) of the European Union, which implicitly promotes the existence of a so-called Right to Explanation of decisions reached by an automated system [20]. The fact that complex AI models inherently lack interpretability due to which a lot of concern was raised among policymakers and practitioners, has directly undermined the concept of accountability and ethics when it comes to the automated system [19]. The XAI techniques are therefore vital to regulatory compliance as they offer the transparency needed to support AI-based decision-making in critical areas [20]. More so, the accountability of a model is preconditioned by its transparency, particularly, when AI systems gain significant power in such areas as criminal justice and finance [24]. The black-box quality of AI systems is a significant bottleneck to its implementation in mission critical application areas like banking, public safety, and more so, healthcare because of a fundamental constraint of trust in its reliability [32]. Failing to comprehend the reasoning of an AI system poses a fatal limitation on the trust of humans when a life-or-death situation in the medical field could be involved [23]. Explainable AI is aimed at reducing this problem by making the processes of decision-making in the model transparent to explain how the system is more reliable and trustworthy overall [20]. The solution to this trust challenge is a critical research topic because the absence of transparency is a direct barrier to the implementation of high-performance AI in areas where the reliability of the result is crucial [24]. Explainability is a crucial part of responsible AI creation, which offers the required transparency to debug models and be fair. The development of the sophisticated AI systems has predisposed them to both security attacks and the introduction of a significant degree of bias into their results [18].

XAI methods, including feature attribution and surrogate modeling [20], can enable the researcher to detect and address these biases and ethical issues, which is necessary to build fair systems [24]. With this perspective,

transparency enables developers and other auditors to identify weaknesses, rectify faults, and minimize the unwanted false negative and false positive results which are likely to emerge in non-transparent black-box models [32]. XAI is inherently connected to the idea of advancing the scientific knowledge by transforming AI as a black box into a generator of knowledge. XAI focuses on the issue of making AI algorithms and the resulting decision comprehensible to humans [23]. Such interpretability, especially in the case of complex models such as Large Language Models (LLM), is obtained through the provision of transparent understandings of the way the model works [20]. XAI will help mitigate the black box by solving the gap between the high-performance AI accuracy and human-understandable explanations [20]. This openness is essential in enhancing a higher predictability, standardizing the assessment of methods and finally developing the theoretical and practical research of AI [21]. Moreover, XAI will promote trust among the users, and they would be more inclined to embrace the AI technologies when they can see the reasoning behind the decision. Trust is essential in such a context such as autonomous vehicles or predictive policing, and these decisions directly affect safety and societal outcomes [25]. Not only is XAI a technical requirement, but also a foundation to the ethical and successful implementation of AI technologies. XAI also focuses on ensuring that AI system is interpretable and accountable so that these systems can be in line with human values and expectations and overcome crucial issues of fairness, transparency, and trust [4].

Large Language Models (LLMs) are deep learning models that are based on the transformer architecture, processing and generating human language, and making use of the strength of self-attention mechanism to learn complex and long-range dependencies in text with weights being computed on the words irrespective of their location. The model begins by encoding the input words (tokens) in the embedding layer to high-dimensional semantic vectors, then through the various layers of the transformer, repeated matrix tasks, the self-attention operation, repeatedly refines the vector representation and calculates the distances among the tokens, finally, allowing the model to generate the correct output. As an example, in the figure 1, the transformer architecture had forecasted the output The, cat, sat, on, the, mat. LLMs have also shown impressive performance in other natural language processing workloads, such as translation, summarization, and question answering and thus form part of the current research and practice in AI [3]; in education, they are being utilized in personalized learning, automatic tutoring, and content generation, providing students with personalized education experiences [9], [10].

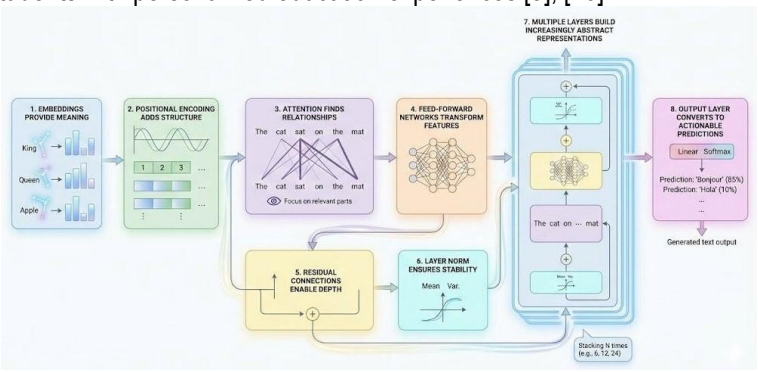


Figure 1: Large Language Models (LLMs) Transformer Architecture

From the figure 1, the steps involved are:

Step 1 - Embeddings Provide Meaning: Words (like "King" or "Apple") are converted into high-dimensional vectors. These vectors ensure that words with similar meanings are mathematically close to one another in a conceptual space.

Step 2 - Positional Encoding Adds Structure: Since Transformers process all words in a sentence simultaneously (rather than one by one), they need a way to know the order of words. Mathematical sine and cosine waves are added to the embeddings to "stamp" each word with its position in the sentence.

Step 3 - Attention Finds Relationships: This is the "magic" of the Transformer. The Self-Attention mechanism allows the model to look at every word in a sentence and determine which other words are most relevant to it (e.g., in the phrase "the cat sat on the mat," the word "sat" focuses heavily on "cat").

Step 4 - Feed-Forward Networks Transform Features: Once the relationships are identified, these neural networks further process each word's representation independently, refining the features and preparing them for the next layer.

Step 5 - Residual Connections Enable Depth: These are "short-circuits" that allow information to skip certain layers. This helps gradients flow through the network during training, preventing the model from forgetting earlier information as it gets deeper.

Step 6 - Layer Norm Ensures Stability: Normalization keeps the mathematical values within a consistent range, which prevents the "exploding" or "vanishing" of numbers that can break the learning process.

Step 7 - Multiple Layers Build Abstract Representations: By stacking these blocks N times, the model moves from understanding simple grammar to understanding complex themes and logic.

Step 8 - Output Layer: The final representation is passed through a Linear layer and a Softmax function. This converts the internal numbers into probabilities for the entire vocabulary.

Large Language Models (LLMs) are deep learning models that are based on the transformer architecture, processing and generating human language, and making use of the strength of self-attention mechanism to learn complex and long-range dependencies in text with weights being computed on the words irrespective of their location. The model begins by encoding the input words (tokens) in the embedding layer to high-dimensional semantic vectors, then through the various layers of the transformer, repeated matrix tasks, the self-attention operation, repeatedly refines the vector representation and calculates the distances among the tokens, finally, allowing the model to generate the correct output. As an example, in the figure 1, the transformer architecture had forecasted the output The, cat, sat, on, the, mat. LLMs have also shown impressive performance in other natural language processing workloads, such as translation, summarization, and question answering and thus form part of the current research and practice in AI [3]; in education, they are being utilized in personalized learning, automatic tutoring, and content generation, providing students with personalized education experiences [9], [10].

Because of the complexity of LLMs, several approaches have been created to shed some light into their decision-making processes, so that they are more interpretable and readable. The approaches have been broadly divided in to Model-Agnostic Post-Hoc Interpreters (MAPHI) [11], Intrinsically Interpretable Models (IIMs) [12] and Overarching Frameworks and Auditing (OFA) [13]. The use of MAPHI technologies follows the prediction of the situation or instances with visualization of attention by the model, especially when it involves a transformer-based model; it indicates the sections of the input text, which the model pays attention to. to make predictions or to bring out the correlations among tokens in a sentence, this resulted in the extent to which the model places value to certain words or phrases in coming up with responses [14]. LIME (Local Interpretable Model-agnostic Explanations) a powerful post-hoc method, a simple interpretable model, which generates a local explanation by perturbing the input and watching the prediction change of the model, and offers insights into the factors that impact the decision [15]. Another post-hoc method that is in the category of SHAP (Shapley Additive Explanations) uses cooperative game theory to assign a contribution score to each feature, the values of which can be used to understand the influence of each token in the input on the model, in both global and local senses [16]. Surrogate model, a post-hoc method is also an intrinsically interpretable model which has been trained on data to act in a way similar to the complex black-box model, such as decision tree model or a linear model. Instead, Intrinsically Interpretable Models (IIMs) attempts to create models that can be more intrinsically interpretable. As an example, decision trees or rule-based systems are more straightforward and their cause and effect can be easily described compared to LLMs.

These models might not be as predictive as LLMs but they offer a better view of how decisions are made. A hybrid approach of these more complex LLMs and simpler models could be to combine both and have high accuracy with clarity offered by the simpler model [17]. Overarching Frameworks and Auditing (OFA) should support the application of XAI techniques to individual black-box models to ensure that the resulting transparency does result in actual accountability and reliable system deployment [18]. General frameworks such as Responsible AI (RAI) and AI Governance give the obligatory framework of this trust that comprises a set of principles such as Fairness, Accountability, Transparency, and Ethics (FATE) that need to be followed during the AI lifecycle [18]. This transforms the notion of transparency to an achievable characteristic to an ethical and legal duty particularly in the areas that are key to a mission like healthcare, education and finance [19].

The governance structure requires the AI systems to be ethical and fair in order to offer the required visibility to detect and reduce the bias, which is concealed in the black-box models, such as discriminatory features weighting thus contributing to the regulatory compliance [20]. OFA uses auditing and Compliance techniques. Auditing is the feasible way through which high-level frameworks are imposed so that the explanations given are not merely plausible, but also faithful, stable and applied in a responsible way when requirements of regulatory compliance require it to be justified as to automated decisions in areas such as finance, education and healthcare. The audit trails required to prove the compliance with these legal and ethical requirements can be the documentation and explanations produced by the XAI tools (feature attributions or counterfactuals), and the model behavior can be justified in a court or to any other regulatory organization [19].

In addition to the ethical governance, a technical framework is required in order to govern the quality of the XAI tools themselves. A good example of an attempt aimed at standardizing the technical audit of explanation methods is the OpenXAI framework [19], which is currently developed to systematically evaluate and benchmark the post-hoc explanation methods (such as LIME and SHAP). It also offers an end-to-end automated pipeline that facilitates standardization and reproducibility in XAI research with the following key metrics; faithfulness, stability (robustness) and fairness [21].

Approaches and Techniques

As stated earlier in this paper, Explainable AI (XAI) is an array of models and techniques that aim to increase the transparency and understandability of AI models and gain a better insight into how the machine arrives at its conclusion and inspire stakeholders to trust AI. These methods in XAI may generally be generalized into post-hoc explainability and intrinsic explainability, which discuss different areas of interpretability [7]. Post-hoc explainability is the approach that is used after training a model. Another technique that is one of the most popular in this category is LIME (Local Interpretable Model-agnostic Explanations), which generates local surrogate models to elucidate individual predictions [26]. The SHAP (Shapley Additive Explanations) is another famous technique that uses feature importance scores that are based on Shapley values, which have their origin in cooperative game theory, SHAP is also useful in education system where the students are helped to make sense of the feedback and the self-regulated learning. SHAP can be used both globally and locally, which is why it is versatile to multiple applications [27]. The visualization techniques are also part of the post-hoc techniques.

In one case, saliency maps and heatmaps are widely applied in the image processing activity to outline regions in an input image that affect a model greatly in its prediction. These visual tools are especially useful to convolutional neural networks (CNNs) [28]. In the case of text-based models, methods like attention mechanisms can be used to show what is the most relevant part of input sequence to a prediction [29]. This works especially well in natural language processing where it can be important to be able to know which words or phrases make a difference. Another post-hoc approach is counterfactual analysis, which considers the impact of minor variations in input data to predictions, which illuminates the decision limits and model behavior. Instead, intrinsic explainability emphasizes the development of models in a manner that is easy to interpret. These would be decision trees, linear regression and rule-based systems, with the decision-making process being understandable in nature [30]. These models are not as complex and in many cases they do not predict as well as deep learning models but they are more favored in those situations where the most important attribute is interpretability.

Another method which can provide explainability is prototype-based techniques which explain using representative examples of the training data. These instances explain how a model classifies the data points so that it simplifies the decision-making process by the users [31]. Combination methods Hybrid methods that combine several explainability methods are finally, becoming popular. As an example, visualization tools and feature importance scores, interpretable surrogate models and attention mechanisms can be used together, which allow gaining a more in-depth insight into model behavior [32]. These fundamental practices and methods of XAI can support various functions, such as technical debugging and optimization as well as ethical or compliance and stakeholder communication because of such practices, AI practitioners can guarantee that their models are not only true but can also be interpreted, held to account and consistent with human expectations [26].

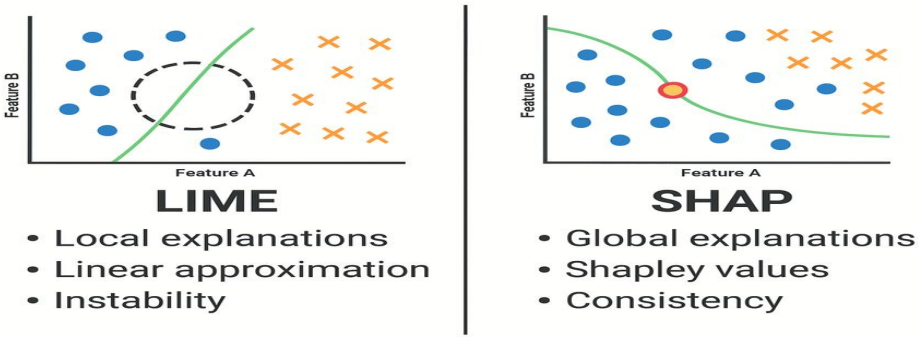


Figure 2: Graphical comparison of LIME and SHAP

Figure 2 is a comparative description of LIME and SHAP, which are two of the most popular techniques employed in eXplainable AI (XAI) to gain insight into complex black-box machine learning models. The visual displays the way that both approaches are trying to solve the issue of explaining a given prediction through examining the connection between the features of input (Feature A and Feature B) and the boundary of the decision of the model.

LIME is based on the concept of local surrogate modeling. The overall distribution of blue dots and orange crosses shows the complex, non-linear decision boundary of the original model as indicated in the left panel. LIME considers a particular case, which is marked by the dashed circle and makes a simplified, linear approximation (the straight green line) that can be considered true only in that small neighborhood. This renders it very useful in making local explanations, the user can get to know why a given decision was made. Nevertheless, due to the fact that the linear model is estimated on a sampled neighborhood, it may be unstable; changing the neighborhood slightly or the sampling procedure may lead to the establishment of an alternative description.

An approach to cooperative game theory, SHAP relies more on mathematically rigorously, namely Shapley values. The green line in the right panel is a more inclined to the holistic perspective of the behavior of the model over the feature space. In contrast to the local snapshot at LIME, SHAP shows how much each feature contributes to the prediction, taking all combinations of features into account, providing explanations of the prediction globally, and being consistent throughout the entire dataset. The red dot on the curve indicates a very particular spot at which the features are being estimated. The first strength of SHAP is consistency; it ensures that in cases where a model is modified, and a feature becomes less significant, the importance value of that feature will not reduce.

Table 1: LIME vs. SHAP Comparison

Method	Type	Key Difference	Strengths	Limitations	Use Case
LIME	Local approximation	Input perturbations for surrogates; rapid but variable	Fast local insights	Less consistent	Rapid individual explanations
SHAP	Shapley values	Game-theoretic fairness/consistency	Accurate, unified attributions	Computationally intensive	Audits, global/regulatory needs

Applications of XAI to Practice

Explainable AI (XAI) is a critical component to the practical applications where the inherent black-box nature of the models restricts trust in critical decisions that can be a matter of life or death or may influence financial stability [21], [24], [19], [23]. XAI is applied in clinical decision support in the medical field, and it is applied to clarify a particular malignancy prediction or diagnosis to clinicians, which results in more precise diagnoses and enables locating tumors locally [21], [23]. XAI is important in finance, including to justify decisions such as the denial of a loan application as required by regulations, which is also the case in the banking industry [21], [19]. In the same manner, XAI transparency is required to be available in the area of a publicly facing service and in autonomous vehicles to be able to comprehend and clarify a particular maneuver [21], [18]. In addition to domain-specific applications, XAI plays an essential role in combating bias and fairness in AI applications, trying to enhance trust, responsibility and fairness offering insights to reduce bias and adhere to regulations [19], [20], [24]. This involves applying interpretability

technologies, like proxy fairness techniques, to audit models with regard to the use of discriminatory features and bias correction [21].

Lastly, XAI is fundamental to the debugging and refining of a model, where explanations are applied to detect the flaws and hence reduce the false negative and false positive results to increase the stability of a model (robustness) and allow the repetition of improvement of the model with increasing prediction accuracy [21], [21]. The introduction of advanced, black-box AI systems in the educational industry raises ethical issues [33], the main ones being algorithmic fairness and the possibility of biasing or discriminating a particular group of users [38], [39]. Explainable AI (XAI) is a vital project aimed at addressing such issues by disillusioning the unknown mechanisms of AI models, offering understandable results of their learning findings [34], [35], [39]. This transparency is the most important in the educational field as it creates trust, a better appreciation of the functioning of AI, and assures that AI-driven tools adhere to high standards of reliability and protection of end-user privacy which is necessary to ensure that both students and educators feel comfortable with the results of the implementation of the system [36], [37], [39].

Moreover, one of the main flaws of various studies on the educational data science is the exclusive emphasis on the maximization of the classification accuracy and a complete disregard of the model interpretation [40]. Nevertheless, to become genuinely efficient in the educational setting, a decision support system should provide more than an efficient prediction; it should also indicate why a particular prediction was undertaken, and, above all, it should suggest the specific intervention that needs to be conducted to help the particular student, which is the local elucidation potential that the XAI offers [40].

Challenges, Evaluation, and Future

Challenges in Evaluation and Standardized Frameworks

1. XAI faces bottlenecks due to the difficulty of objectively quantifying explanation quality and the lack of systematic benchmarking for post-hoc explanations.
2. The OpenXAI framework addresses these issues by emphasizing standardized, quantitative metrics to automate the evaluation process.
3. OpenXAI introduces twenty-two quantitative metrics across three technical dimensions:
    a. Faithfulness: The accuracy of the explanation relative to the model's actual decision-making.
    b. Stability: How susceptible an explanation is to small changes in input data.
    c. Fairness: The consistency of explanations across different demographic or input groups.
4. Technically sound explanations are a prerequisite for human trust, as unfaithful or unstable explanations are practically useless to the end-user.
5. Quantitative methods, including end-to-end pipelines and public leaderboards, promote transparency and reproducibility in the field.

The Accuracy-Explainability Trade-off

1. A "Trade-off Dilemma" exists in AI, particularly with the growth of complex systems like Deep Learning (DL) and Large Language Models (LLMs).
2. Current high-performance systems often rely on non-linear structures that are inherently difficult to interpret, leading to a lack of transparency.
3. This opacity limits AI implementation in mission-critical areas (e.g., healthcare and finance) where trust and ethical accountability are paramount.
4. XAI aims to bridge the gap between high performance and human-understandable explanations without sacrificing predictive accuracy.

Security and Ethical Risks

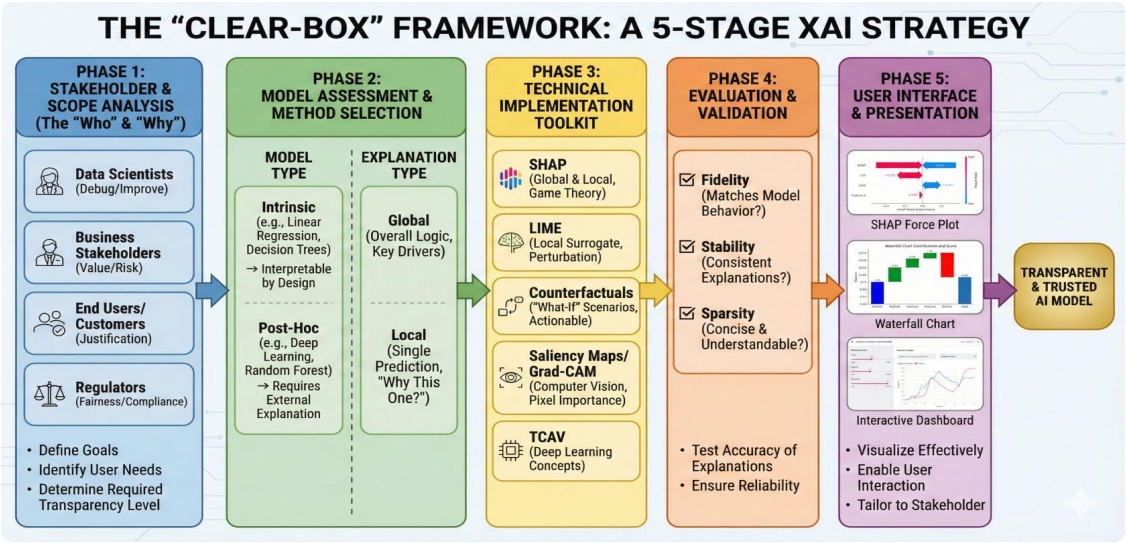
1. XAI systems must be resistant to malicious manipulation to maintain public trust.
2. Explanation hacking occurs when users deliberately force an AI to provide convincing but incorrect explanations of its decisions.
3. "Explanation washing" is a major threat where technical explanations are used to mask underlying biases or sanction unfair results.

4. Fabricated explanations can justify discriminatory outcomes, such as biased loan rejections, which perpetuates systemic inequalities.
5. Future research must focus on developing detection and prevention mechanisms to ensure XAI remains a tool for transparency rather than obfuscation.
6. Interactive and Human-in-the-Loop XAI
7. Involving the user in generating and refining explanations is crucial, as static explanations are often insufficient for complex models.
8. This "humanistic approach" transforms the process into a dialogue, allowing users to ask active questions (e.g., why, why not, and what-if).
9. Counterfactual and contrastive explanations are frequently used to provide practical insights:
    a. "Why" questions seek justification for a specific outcome.
    b. "Why not" questions use counterfactuals to see if changing an input would alter the prediction.
10. The iterative feedback loop helps build confidence, increase behavioral knowledge, and mitigate model vulnerabilities in critical environments.
11. The ultimate goal is a personalized, dynamic experience that meets the specific informational needs of the end-user.

The Clear-Box XAI Implementation Framework

Figure 3: The "Clear-Box" Framework: A 5-Stage XAI Strategy

Phase 1:



Stakeholder and Scope Analysis (The Who and the Why)

This step brought out the consumer of the explanation. There are various users who would need various degrees of transparency. Data Scientists/Engineers: Require technical specifications to debug the model, optimize it, and identify overfitting that will drive the emphasis on gradient flows, layer activations, detailed feature importance. Business Stakeholders: They should evaluate value, risk, and alignment to business logic that will cause them to pay attention to global model behavior which serves as major drivers of forecasts. End Users/Customers: Requirement of justification of a certain decision (e.g., "Why was my loan denied?") will make them target local explanations, actionable counterfactuals. Regulators: should guarantee fairness, non-discrimination and compliance (e.g., GDPR, CCPA) that will make them concentrate on detecting bias, fairness metrics, robustness.

Phase 2: Model Testing and Methodology

This is where the model character is decided and the kind of explanation is to be given. Intrinsic vs. Post-Hoc Intrinsic (Interpretable by Design): When the sole measure is not high-accuracy, then abandon black-box models in favor of transparent models such as Linear Regression, Decision Trees, or GAMs (Generalized Additive Models). Post-Hoc (After-the-fact): When you have to use a complex model (e.g., Random Forest, Deep Learning), you have to explain it using other methods. Explanations Global vs. Local. Global Interpretability: Gives the general reasoning of the model. As an illustration, "What are the characteristics of the model that the model typically considers to be most important? Local Interpretability: Interprets an individual prediction. As an example, why have the model identified 80% churn risk on this particular customer?

Phase 3: Technical Implementation Toolkit

This step defines the actual algorithms to be adopted in accordance to the choices in Phase 2. Technique Type Best Used for SHAP (SHapley Additive exPlanations) Local & Global The Standard of Gold. According to the game theory; gives the same value in contributions of features. Good for structured data. LIME (Local Interpretable Model-agnostic Explanations) Local A perturbation of data around a single prediction to train a simple surrogate model. Good when one wants to see what happens. Counterfactuals Local User-centric. There would have been an increase in loan approval had you earned an extra 5k. Actionable advice. Saliency Maps / Grad-CAM Local Computer Vision. Shows the pixels in an image that made the biggest contribution to the classification. TCAV (Testing with Concept Activation Vectors) Global Deep Learning. Represent predictions in high level terms (e.g., Stripes) and not in pixel terms. Technical Note on SHAP: The marginal contribution of feature i in any combination of features of a dataset is then taken as the average of all coalitions of features, which gives a feature value F! of the feature.

Phase 4: Evaluation and validation

The phase is utilized to prove the XAI methods. Fidelity:Is the explanation consistent with the behavior of the model? (When the explanation indicates that Feature A matters, then when Feature A is removed the prediction should alter dramatically). Stability: Does the explanation change with an insignificant change in the input (even though the prediction does not)? (Unstable explanations are undermining trust). Sparsity: Does the explanation take a long time to explain? (A list of 5 major features would work better than 100).

Phase 5 User Interface and Presentation

This stage showed the visualization of the data. Interpretation SHAP/LIME: The use of force plots or waterfall charts can be used to visualize how features drive the prediction up (positive force) or down (negative force) of the prediction baseline. Interactive Dashboards: Construct tools (with Streamlit or Dash) so that stakeholders can interactively play (through feature sliders) with the model and explanation and see the response of the model and the explanation.

Results and Discussion

Based on the research, a 5-stage Clear-Box framework is developed that operationalizes transparency in opaque AI systems, namely, the challenge of interpreting Deep Learning and Large Language Models (LLM) with the help of a systematic approach. The findings suggest a successful XAI implementation should be based on a "Stakeholder & Scope Analysis" to customize explanations which can give technical debugging information to data scientists and justify and counterfactuals to end-users and regulators to address the tension between information requirements of different users. It is mentioned in the discussion that trading-off of between model accuracy and interpretability must

be done strategically between Intrinsically Interpretable Models (IIMs) and Model-Agnostic Post-Hoc Interpreters (MAPHI) such as LIME and SHAP the latter with a Gold Standard consistency through cooperative game theory. Moreover, the framework points out that to reduce risks like the so-called explanation hacking and guarantee compliance with the regulations (e.g., GDPR), XAI tools need to be introduced to strict evaluation stages that address the fidelity, stability, and sparsity, and ultimately presented in interactive dashboards to contribute to the development of the actual trust of the user.

Future Directions and Policy Implications

According to the challenges and future directions of the study, the following recommendations are suggested in the successful implementation of XAI: Implement Standardized Evaluation Frameworks: To break the bottleneck of objectively measuring the quality of explanation, organizations are advised to implement such systems as OpenXAI. They should be evaluated in terms of quantitative measures such as: Faithfulness: Making sure that the description is true to the process of the decision model. Stability: This is to make sure that explanations are stable even when inputs change slightly. Fairness: This is the consistency within various demographic groups. Adopt Overarching Governance (OFA): Overarching Frameworks and Auditing (OFA) must be able to support technical XAI application. Responsible AI (RAI) principles and Fairness, Accountability, Transparency, and Ethics (FATE) should be enforced in the framework of the AI lifecycle to make systems trustworthy and compliant. Create Human-in-the-Loop Interactive Systems: Static explanations can be too little. It has been suggested to construct Interactive XAI in which users can query the model using the questions, why, why not, and what-if queries.

This method will turn explanation into a conversation, which will develop more understanding and trust, particularly in areas of mission critical areas. Reduce "Explanation Hacking": The creators should be careful of malicious actors who may abuse explanations to obscure biases, or practice explanation washing. In the future, the development should be based on a sound system of detection to make sure that explanations are not a tool of obfuscation but a tool of transparency. Apply Hybrid Modeling Approaches: Use a simple model where possible and complex LLMs. This mixed methodology enables the LLM to give high-accuracy and the less complex model to give clarity, which is an optimal trade-off between the performance and the interpretability. Explain to Stakeholders: This assists in delivering the developers insights on how to behave globally, and deliver counterfactuals to the end-users to act on. 6.0 Conclusion The development of complex, black-box AI systems, especially Deep Learning (DL) and Large Language Models (LLMs) has generated an urgent necessity to establish Explainable Artificial Intelligence (XAI) to enforce transparency, trust, and accountability in the high-stakes sectors such as medicine, finance, and education. XAI attempts to solve the natural opaqueness of these models by offering methods, such as the Model-Agnostic Post-Hoc Interpreters (MAPHI) such as LIME and SHAP, and Intrinsically Interpretable Models (IIMs) that explain decision making, and as a result connecting high-performance accuracy to human interpretability.

The significance of XAI is not limited to technical competencies, but rather to satisfying the high standards of the Overarching Frameworks and Auditing (OFA) such as the ethical and legal requirement of Responsible AI (RAI) and regulation such as GDPR, which requires the need to justify the automated decision-making. Future directions include standardized, quantitative assessment based on models such as OpenXAI to assess faithfulness, stability, and fairness and also the development of Interactive and Human-in-the-Loop XAI to offer dynamic and user-friendly explanations that develop real trust and resistance to adversaries such as explanation hacking.

Conclusion

Overall, although the recent swift development of advanced AI systems such as Deep Learning and Large Language Models can be unprecedentedly efficient, it introduces a so-called Trade-off Dilemma, as the nature of such systems makes them intrinsically opaque and thus undermines trust and responsibility in missions-critical applications. Explainable Artificial Intelligence (XAI) can play the crucial role in this mismatch, going beyond technical enlightenment as models are required to be faithful, stable, and equitable by uniform systems such as OpenXAI. Nevertheless, the field should stay alert on the ethical hazards like the so-called explanation washing and wicked manipulation by focusing on humanistic and interactive conversations that allow users to interact with models and obtain why and what-if answers. Finally, one can conclude that the future potential of AI use is based on the creation of strict, open-source evaluation pipelines that can convert XAI into a moving device rather than a fixed object of human and machine cooperation.

References

[1] A. Vaswani et al., "Attention is all you need," Adv. Neural Inf. Process. Syst., 2017.

[2] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," IEEE Access, vol. 6, pp. 52138–52160, 2018.

[3] C. Molnar, Interpretable Machine Learning, 2nd ed. 2022.

[4] D. Gunning, "Explainable artificial intelligence (XAI)," DARPA, 2017.

[5] A. B. Arrieta et al., "Explainable AI (XAI): Concepts, taxonomies...," Inf. Fusion, vol. 58, pp. 82–115, 2020.

[6] T. Miller, "Explanation in AI: Insights from social sciences," Artif. Intell., vol. 267, pp. 1–38, 2019.

[7] EU GDPR, Regulation (EU) 2016/679, 2016.

[8] G. El-Khawaga, "Evaluating XAI methods in the context of predictive process monitoring," Ph.D. dissertation, University of Ulm, Germany, 2024.

[9] M. D. Idris, X. Feng, and V. Dyo, "Revolutionizing higher education: Unleashing the potential of large language models for strategic transformation," *IEEE Access*, vol. 12, pp. 67 738–67 757, 2024.

[10] Baker, R. S., & Siemens, G. (2022). Learning analytics and educational data mining. In R. K. Sawyer (Ed.), *The Cambridge Handbook of the Learning Sciences* (3rd ed., pp. 259–278). Cambridge University Press.

[11] Ouyang, F., & Jiao, P. (2021). Artificial intelligence in education: The three paradigms. *Computers and Education: Artificial Intelligence*, 2, 100020.

[12] M. Miró-Nicolau, A. Jaume-i-Capó, and G. Moyà-Alcover, "A comprehensive study on fidelity metrics for XAI," *Inf. Process. Manag.*, vol. 62, no. 1, p. 103900, 2025.

[13] S. Krishna, J. Ma, D. Slack, A. Ghandeharioun, S. Singh, and H. Lakkaraju, "Post hoc explanations of language models can improve language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[14] A. Agiollo, L. C. Siebert, P. K. Murukannaiah, and A. Omicini, "From large language models to small logic programs: building global explanations from disagreeing local post-hoc explainers," *Auton. Agents Multi Agent Syst.*, vol. 38, no. 2, p. 32, 2024.

[15] S. Luo, H. Ivison, S. C. Han, and J. Poon, "Local interpretations for explainable natural language processing: A survey," *ACM Computing Surveys*, vol. 56, no. 9, pp. 1–36, 2024.

[16] Q. Zhang, Y. Hu, J. Yan, H. Zhang, X. Xie, J. Zhu, H. Li, X. Niu, L. Li, Y. Sun et al., "Large-language-model-based ai agent for organic semiconductor device research," *Advanced Materials*, vol. 36, no. 32, p. 2405163, 2024.

[17] Y. Ivanov, "Understanding the inner workings of large language models: Interpretability and explainability," *MZ Journal of Artificial Intelligence*, vol. 1, no. 1, pp. 1–5, 2024.

[18] V. Chamola et al., "A Review of Trustworthy and Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 11, pp. 78985–79015, 2023. (Note: The page ranges vary slightly, but the citation is the same core article).

[19] S. Akhai, "From Black Boxes to Transparent Machines: The Quest for Explainable AI," *SSRN Electronic Journal*, 2023.

[20] M. E. C. Souza and L. Weigang, "Unveiling the Black Box: The Significance of XAI in Making LLMs Transparent," *Inteligencia Artificial*, 2025.

[21] C. Agarwal et al., "OpenXAI: Towards a Transparent Evaluation of Post hoc Model Explanations," in *NeurIPS*, 2022.

[22] V. Hassija et al., "Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence," *Cognitive Comput.*, vol. 16, pp. 45–74, 2024.

[23] T. Hulsen, "Explainable Artificial Intelligence (XAI): Concepts and Challenges in Healthcare," *AI*, 202X.

[24] N. Patidar et al., "Transparency in AI Decision Making: A Survey of Explainable AI Methods and Applications," *Advances of Robotic Technology*, vol. 2, no. 1, 2024.

[25] S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel, "Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions," *IEEE Access*, 2024.

[26] C. O. Retzlaff, A. Angerschmid, A. Saranti, D. Schneeberger, R. Röttger, H. Müller, and A. Holzinger, “Post-hoc vs antehoc explanations: xai design guidelines for data scientists,” *Cogn. Syst. Res.*, vol. 86, p. 101243, 2024.

[27] T. Clement, T. T. H. Nguyen, N. Kemmerzell, M. Abdelaal, and D. Stjelja, “Beyond explaining: Xai-based adaptive learning with SHAP clustering for energy consumption prediction,” *CoRR*, vol. abs/2402.04982, 2024.

[28] H. Kim, J. Jung, R. Hwang, S. Park, S. Lee, G. Kim, and B. W. Lee, “Classification of PRPD pattern in cast-resin transformers using CNN and implementation of explainable AI (XAI) with grad-cam,”*IEEE Access*, vol. 12, pp. 53 623–53 632, 2024.

[29] E. H. Zaryabi, L. Moradi, B. Kalantar, N. Ueda, and A. A. Halin, “Unboxing the black box of attention mechanisms in remote sensing big data using XAI,” *Remote. Sens.*, vol. 14, no. 24, p. 6254, 2022.

[30] J.-X. Mi, X. Jiang, L. Luo, and Y. Gao, “Toward explainable artificial intelligence: A survey and overview on their intrinsic properties,” *Neurocomputing*, vol. 563, p. 126919, 2024.

[31] S. Pathak, J. Schlötterer, J. Veltman, J. Geerdink, M. van Keulen, and C. Seifert, “Prototype-based interpretable breast cancer prediction models: Analysis and challenges,” in *Explainable Artificial Intelligence - Second World Conference, xAI 2024*, 2024, pp. 21–42.

[32] B. Biswas, A. Mukhopadhyay, A. Kumar, and D. Delen, “A hybrid framework using explainable AI (XAI) in cyber-risk management for defence and recovery against phishing attacks,” *Decis. Support Syst.*, vol. 177, p. 114102, 2024.

[33] T. F. Tahiru, "AI in education: A systematic literature review," *Journal of Cases on Information Technology (JCIT)*, vol. 23, no. 1, pp. 1-20, Jan. 2021.

[34] S. Galhotra, R. Addanki, and B. Saha, "How to Design Robust Algorithms using Noisy Comparison Oracle," *Proceedings of the VLDB Endowment*, vol. 14, no. 10, pp. 1703–1716, 2021.

[35] A. Rai, "Explainable AI: From black box to glass box," *Journal of the Academy of Marketing Science*, vol. 48, pp. 137-141, 2020.

[36] A. Hanif, X. Zhang, and S. Wood, "A survey on explainable artificial intelligence techniques and challenges," *2021 IEEE 25th International Enterprise Distributed Object Computing Workshop (EDOCW)*, 2021, pp. 81-89.

[37] T. Hasib, et al., "A Machine Learning and Explainable AI Approach for Predicting Secondary School Student Performance," *Proceedings of the 2022 International Conference on Information Systems (ICIS)*, 2022.

[38] H. Kaur, et al., "Artificial Intelligence for Cybersecurity: Literature Review and Future Research Directions," *Sensors*, vol. 22, no. 1, p. 306, 2022.

[39] Z. M. Altukhi and S. Pradhan, "Systematic Literature Review: Explainable AI Definitions and Challenges in Education," *Forty-Fifth International Conference on Information Systems (ICIS)*, 2024.

[40] M. Nagy and R. Molontay, "Interpretable Dropout Prediction: Towards XAI-Based Personalized Intervention," *International Journal of Artificial Intelligence in Education*, vol. 34, pp. 274–300, 2024.

[41] P. N. Srinivasu, N. Sandhya, R. H. Jhaveri, and R. Raut, "From Blackbox to Explainable AI in Healthcare: Existing Tools and Case Studies," Mobile Information Systems, vol. 2022, Article ID 8167821, 2022. (Sources: uploaded:Mobile Information Systems - 2022 - Srinivasu - From Blackbox to Explainable AI in Healthcare Existing Tools and Case.pdf)

[42] J. Jiao, S. Afroogh, Y. Xu, and C. Phillips, “Is a picture worth a thousand explanations? adversarial helpfulness of black-box explanations can be adversarially helpful,” arXiv preprint arXiv:2405.06800, 2024.

[43] J. Jiao, S. Afroogh, Y. Xu, and C. Phillips, “Navigating llm ethics: Advancements, challenges, and future directions,” arXiv preprint arXiv:2406.18841, 2024.

[44] S. S. Adadi and M. Berrada, “Peeking inside the black-box: a survey on explainable artificial intelligence (XAI),” IEEE Access, vol. [cite_start]6, pp. 52138–52161, 2018.

[45] D. T. Gunning, “Explainable artificial intelligence (XAI),” Defense Adv. Res. Proj. [cite_start]Agency (DARPA), 2017.

[46] J. B. van der Waa, J. van Diggelen, F. E. J. C. Peeters, and H. van der Velden, “Contrastive explanations for reinforcement learning in terms of expected reward deviations,” Artif. Intell., vol. 297, p. [cite_start]103525, 2021.

[47] R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, and F. Giannotti, “A unified approach for explaining the predictions of any classifier,” IEEE Trans. Knowl. Data Eng., vol. 34, no. [cite_start]1, pp. 248–261, 2022.

[48] R. R. A. M. Islam, S. N. Akter, R. F. I. Arafat, S. H. Sany, M. A. Al Maruf, and M. I. Hossen, “Explainable artificial intelligence (XAI) in image processing: A systematic review,” SN Comput. Sci., vol. 4, no. 1, p. [cite_start]10, 2023.