Review

Enhancing AI Transparency for Human Understanding: A Comprehensive Review

Malika Ara *

Posted Date: 3 October 2024

doi: 10.20944/preprints202410.0262.v1

Keywords: artificial intelligence; machine learning; blackbox; explainable artificial intelligence

 doi:10.20944/preprints202410.0262.v1

Article

Enhancing AI Transparency for Human Understanding: A Comprehensive Review

Malika Ara

Cyber Security, Saint Louis University, Missouri, USA; malika.ara@slu.edu

Abstract: One of the most hotly debated topics in technology is the transparency between AI models and humans. As artificial intelligence (AI) continues to permeate various sectors, the demand for transparency in AI decision-making has become increasingly critical. This paper presents a comprehensive review of Explainable Artificial Intelligence (XAI), examining 57 key studies that focusing on various explanation approaches and their impact on the trust and accountability of end-users. Recognizing the obstacles resulting from the black-box nature of AI models, this work focuses on the need for the proper methods that can be used in the explanation process, enabling both people and AI models to work together. The findings highlight the importance of XAI in enhancing trust, particularly in complex environments such as healthcare and finance, and propose directions for future research to further develop reliable and interpretable AI solutions.

Keywords: artificial intelligence; machine learning; blackbox; explainable artificial intelligence

I. Introduction

The growing industry of AI has brought transparency to the centre of the technological conversation. Artificial Intelligence (AI) is the science and engineering of making intelligent machines [1] that can simulate human intelligence to solve real-world problems. The AI system/model is trained with the use of machine learning and deep learning algorithms using the data to think and learn like a human. Even with all this training, there is always a doubt about the lack of transparency and interpretability in these models. Allowing AI to explain itself will allow the humans to trust the solutions and answers given and understand the logic behind it. AI must be trusted to function reliably, trusted to be able to explain conclusions, trusted not to violate privacy, and trusted not to exhibit socially harmful bias [2]. Fig. 1 explains about what AI needs to earn our confidence. Surrounding this central theme are key factors such as responsibility, fairness, transparency, interactivity, stability, satisfaction, robustness, integrability, and explainability. These interconnected elements collectively contribute to a trustworthy AI, ensuring that it aligns with human values, operates reliably, and is easy to understand and interact with. By addressing these factors, AI can build a foundation of trust, enabling its widespread adoption and positive impact on society.

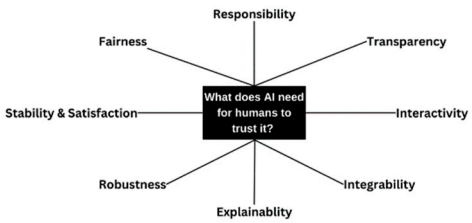

Figure 1. Key factors for human trust in AI.

Several types of AI, such as analytical, functional, interactive, textual, and visual, can be applied to enhance the capabilities of applications.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2024 doi:10.20944/preprints202410.0262.v1

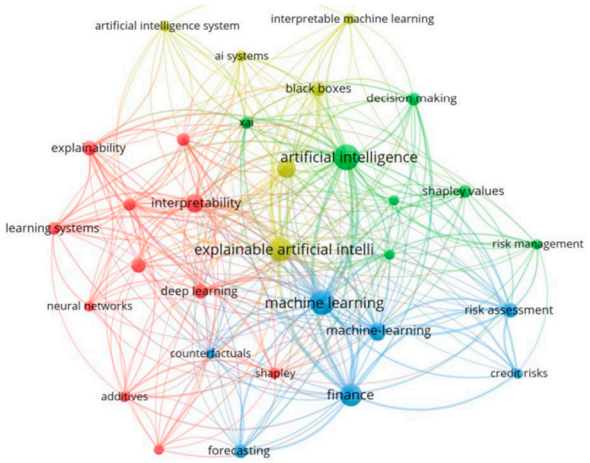

Figure 2. Co-occurrence analysis of the keywords in literature [20].

There is a very slight distinction between explainability and interpretability. Interpretability is the ability to understand a model's internal workings, either in its entirety or in certain areas. In contrast, explainability requires the model to provide justification for its predictions. For example, an AI model that provides a solution will either analyze all the input data or only some portions of it. It does this by figuring out which pieces of information are important to the model and analysing if any changes to the output will be done if there are changes in the information [36].

II. History of AI

The concept of artificial intelligence began to take shape long before John McCarthy coined the term in 1956, with contributions from Vannevar Bush and Alan Turing exploring the potential for machines to simulate human thinking and intelligence. McCarthy's 1956 conference at Dartmouth College marked a pivotal moment, catalysing ongoing efforts to develop intelligent computer systems based on logic and problem-solving, setting the stage for subsequent advancements in AI applications [1].

Many questions arose when artificial intelligence first developed. The need for explanation from the computer models about their solutions or predictions is not a recent question for research. This has been a question for more than 45 years now [3]. From the time the research began in AI, the experts and researchers have argued that the computer systems and AI should be able to prove the steps on which their decision is based. For instance, the neural network model trained on a large dataset of images can recognize objects with high accuracy as shown in fig 3 where the users are unaware of the inner workings (hidden layer) of neural network.

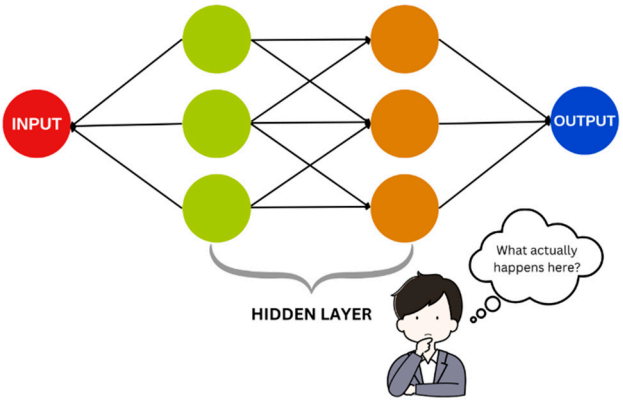

Figure 3. A Simple Neural Network.

 doi:10.20944/preprints202410.0262.v1

Advanced machine learning models, especially deep neural networks, include multiple layers of interconnected nodes (neurons) processing vast amounts of data. The intricate interactions and the sheer scale of computations make it arduous to trace back and articulate how specific decisions are made [5]. Even while it may seem difficult to understand these relationships in their entirety, several frameworks and techniques are developing in response to this challenge.

III. XAI and Its Need

Explainable AI (XAI) is one such strategy for building confidence in AI. Though the need of explaining the process of AI is an old research topic, it has gained popularity as the use of AI systems becomes increasingly prevalent in various domains, such as healthcare, finance, and autonomous vehicles. The goal of XAI is to make the inner workings of AI systems transparent and explainable to humans, to build trust and understanding in AI decision making [4]. This can be done by developing algorithms and models that can be easily understood and interpreted by humans. H. Vainio-Pekka et al. [7] have analysed a total of 142 primary studies in the field of AI ethics emphasizing the importance of XAI in many critical sectors where AI systems directly affect people's physical conditions and safety. The healthcare [19], finance [20], automobiles [21], academia [22], transportation [23], BFSI [24], judicial system [25], insurance [26], manufacturers [27], human resources [28], and defence [29] are some of the essential sectors. IBM Watson, Google DeepMind, Path A, PayPal, Zest Finance, and BlackBox have already begun to use XAI in their respective areas [57]. The need of XAI is underscored by the requirement for transparent ML models to uphold ethical principles such as fairness, accountability, and responsibility.

A. Personalized Perspectives on AI Explanation:

The mind-set of each human being differs from one another, so the level of understanding is different too. The AI system should be able to understand its user and explain its process according to individual needs. [8] introduced four principles of explainable AI based on each users' specific needs and requirements. For example, as shown in Fig 4, the users can be AI developers, policy makers, or end users, that may have varying explanation needs and preferences.

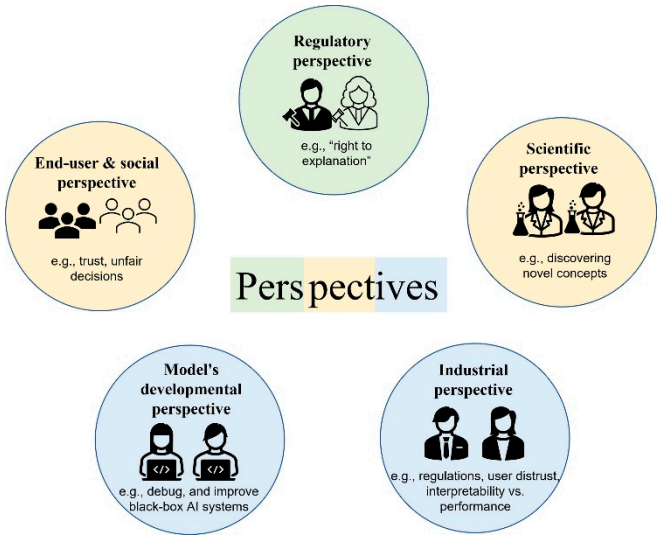

Figure 4. The five main perspectives for the need for XAI. [9].

B. The Interface between AI and Humans:

The interaction between humans and AI is a complex, evolving process that has significant implications for society. It is expected that a successful human-AI team will have collaborative performance that exceeds the performance of humans or AI alone. For a team of human and AI to work it is particularly important to have a mutual understanding of each other's thoughts and observations. [45] and [10] both highlight the need for a deeper understanding of this interaction.

 doi:10.20944/preprints202410.0262.v1

Wang [10] proposed a new research paradigm called the human-AI learning to achieve collaborative performance that surpasses individual capabilities. This study states that the AI should be able to learn from the human experiences and vice versa. This is possible only when the AI architecture and solutions are understandable to humans in a simple human language. To make AI models more transparent and understandable to humans, one approach is to incorporate human knowledge into the training process through methods like knowledge graphs, rule-based systems, and expert collaboration. This can enhance interpretability and reduce the complexity of AI models. Additionally, providing visualizations such as feature importance can help users understand how the AI model makes decisions, increasing transparency and trust in the system.

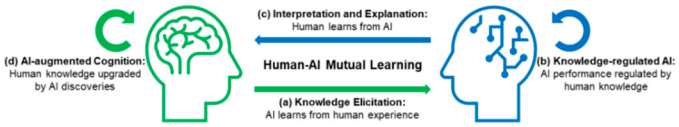

Figure 5. Human-AI mutual learning through knowledge elicitation and knowledge-regulated AI [10].

IV. Top Studies about Transparency in AI

This section covers a range of studies that have explored the field of Explainable AI (XAI) from different perspectives.

A. Study one

This study [11] explored the concept of transparency in artificial intelligence (AI) systems from philosophical, legal, and technological viewpoints. The authors focused on the implementation of transparency across different disciplines in the General Data Protection Regulation (GDPR). They examined eleven cases from which three research were primarily highlighted due to their focus on various aspects of transparency outcomes in various contexts. Ethnographic analysis [46] was conducted to better understand how transparency is perceived and used by companies and users. It is clear from learning about these studies that they all emphasize the value of ethical and accountable concerns in transparency policies across a range of industries:

1. Chen and Sundar [12] approached an experimental method using an eco-friendly mobile app in which transparency played a pivotal role in building the trust among users. They found that the clear and transparent communication between the device/app's friendly features and data usage polices increased the user trust and satisfaction.
2. Moving on to the online advertising sector, [13] conducted a qualitative user study that showed the importance of providing necessary explanations in online ads to build trust and satisfaction among consumers. By transparently communicating the motive behind targeted ads, companies were able to build trust with their audience and improve overall user experience.
3. Social media is such a platform where everyone presents their life in front of everyone. So, it is important for each platform to be trustworthy and transparent. [14] explored the impact of transparency on user awareness within the social media platforms. Their study revealed that transparent communication about algorithmic processes and content curation led to increased user awareness and understanding of platform functionalities.

B. Study two

Zhang [15] and their team's main questions were focused on the following aspects of AI explanations based on the human-AI teaming perspective: 1) publishing ultramodern research, 2) understanding the interaction between trust and effectiveness while exploring personal characteristics. To answer this important question, a carefully constructed mixed factorial survey experiment [16] was conducted, utilizing a between-subjects treatment and two within-subjects' factors [47]. The intervention process involved using scenarios in a simulation game, where the participants had to interact with real-life situations and make judgments of whether the facts provided by AI should be explanatory. Using a predetermined matching procedure, participants were randomly switched to different conditions depending on the human vs. AI partner and whether

 doi:10.20944/preprints202410.0262.v1

they received explanations or not. After watching the feedback videos, the participants underwent an elaborate survey measure that aimed at measuring their perceptions and response towards the treatment and post-task demographic questionnaire.

The results of such a strict method demonstrated compelling patterns, stating the apparent interaction between teammate identification and justification on confidence in the collaboration between humans and AI. The study also demonstrated that explanation design is crucial for promoting trust, coordination, and productivity of the human-AI teamwork and offered insights into structural models for accepting or rejecting the explanation for human and AI teammate groups.

C. Study three

The use of virtual treatment in the metaverse, facilitated by AR and VR glasses, has the potential to revolutionize healthcare delivery [17]. The word "Metaverse" represents a virtual shared area that is a combination of physical and digital worlds where the users can interact with one another. Ali et al. [18] discusses about creating an architecture consisting of three environments namely, healthcare professionals, patients, and AI to provide better and fast healthcare facilities with a realistic experience.

Avatars are the components that are used to communicate within the architecture initiating consultations via voice recordings and the usage of NLP for data extraction. Crucially, the system allows to include the important medical data such as the CT scans and MRI Scans stored securely in the patient's data for the physician's access. The AI models are pre-trained to provide the prediction about the patients' health concerns. But the question is, will the patients be able to trust the AI predictions? For building the trust and making them completely transparent about the conclusions from the AI models, they used GradCAM and LIME approaches to provide logical reasoning to their solutions.

Furthermore, two algorithms were introduced to enhance the functionality and transparency in metaverse based architecture.

Algorithm 1: Artificial-intelligence-based model for the healthcare metaverse

Input: NLP data, Sensor data, Image data, Trained model

Output: O(XAI)

```
1. Procedure AI_model (NLP data, Sensor data, Image data, Trained model)
2.     Data←load_data (NLP_data, sensor_data, Image_data)
3. Data←Data_preprocessing (Data)
4. Prediction←Trained_model.predict(Data)
5. If (Data include medical images) then
       Import GradCAM
       Heatmap=make_gradcam_heatmap(Trained_model, Data)
       Else: Import LIME
       e = LIME.GradientExplainer(Trained_model, Data)
       LIME_values = e.LIME_values(Data)
6. Display_gradcam/LIME
7. O(XAI)←LIME.image_plot(LIME_values)
       O(XAI)←display_gradcam(heatmap)
   8. End.
```

Algorithm 1 addresses the black box nature of the AI Model by evaluating the data and generating the predictions using the XAI techniques (GradCAM and LIME). The insights that are gained from this are used by the medical professionals to take decisions and perform the complex procedures.

 doi:10.20944/preprints202410.0262.v1

Algorithm 2:

Input: Patient P, Doctor D, Consultation query Cq, Caregiver Cg; and Prediction output O(XAI).

Output: Initiate treatment in metaverse IDVEnv (Virtual Environment) and activate patient environment

1. Procedure: Blockchain_metahealth()
2. For Cq by P do
   Response←Fetch_suitable(D)
   If (IDP∈BCP) then
   Assign Caregiver (BCCg)
   Else
   Display ("Patient does not exist, please register")
   End If
   End for
3. Execute_Contract (IDP, tnp, Sig(P))
   Execute_Contract (IDCg, tncg, Sig (Cg))
4. If Dk←Sig (D) then
   Execute−Contract (IDD, tnd)
   Setup VEnv←(IDP, IDD, IDCg, IDVEnv)
   Execute−Contract (IDVEnv)
   Main BC←O(XAI)
   Order Tn on Main BC (IDp, tnp, T, Sig(P)←Verify (Sig (D), Tn, D)
   Order Tn on Main BC (IDp, tnp, T, Sig(P)←Verify (Sig (cg), Tn, Cg)
   End If
5. End

Note: IDP denotes Id of patient, BCP: blockchain of patient, BCCg: blockchain of caregiver, IDCg: Id of caregiver, tncg: transaction pay load of caregiver, IDD: Id of doctor, IDVEnv: Id of virtual environment.

Algorithm 2 is about the information flow within the block chain integrated environment shown in fig 6. When a patient requests a consultation, the system identifies a suitable doctor and initiates the registration process if necessary. A caregiver is assigned, and smart contracts for both patient and caregiver are executed. Upon the doctor's approval via a signed request, their smart contract activates, and treatment begins in a designated virtual environment (IDVEnv). After treatment, the XAI-generated output is stored securely on respective block chains, and the environment is dismantled, ensuring data integrity and security.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2024 doi:10.20944/preprints202410.0262.v1

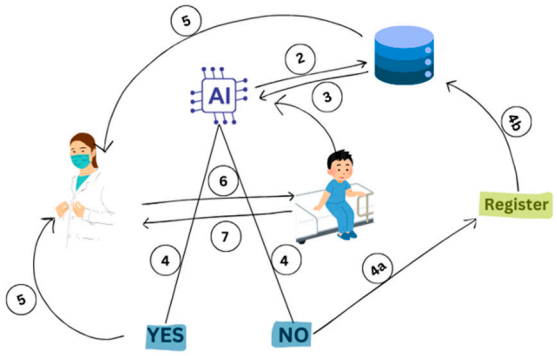

Figure 6. AI registration process (Algorithm 2).

V. Complications and Practical Solutions

A. Problems in today's research

The study [21] clearly explained about the misconceptions on the trustworthiness of AI. XAI is an ever-growing community with lot of technologies coming up in line. The new AI models for example ChatGPT and Gemini are the latest ai models that are used by many individuals today. The answers given by the models are 75% of the time correct, while 25% are wrong. The word strawberry created a controversy in 2024 for not correctly answering the actual number of 'r's present in it. When ChatGPT is asked about the number of 'r''s present in the word 'strawberry' it replies with 2 (fig 7), which is incorrect. On the other hand, when the question is paraphrased and asked it gives the correct answer. These minor differences made by the ai models depreciate the user's trust.

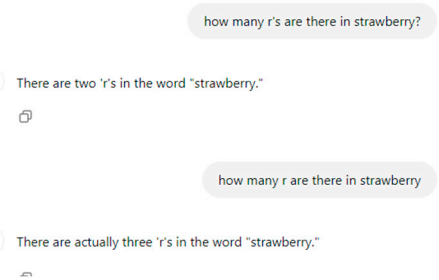

Figure 7. Example representation of answers by AI Model.

There is no explanation for this solution and the user stays confused. This is the black box nature of this model. Black box model means an AI model that is not completely transparent or easily interpretable which makes it hard to understand. To understand learn and explain about a deep neural network, the user would need access to a ton of numbers and there no conceivable way to understand the AI model completely [30].

Figure 8. Use Cases of XAI [48].

B. Can we change these problems to solutions?

Yes, of course! There are various methods and solutions to make AI transparent, but the most effective one is the visualisation of the data and process that goes behind it, that provides an idea and

 doi:10.20944/preprints202410.0262.v1

insight into the dataset. Visualization platforms help users to interactively explore model decisions, offering a more intuitive grasp of how specific input features lead to specific outputs [6]. Visualization and data analysis are the most important tools for getting out the meaningful information for a long time.

1. *Challenge 1 Computation Intensity:* Large datasets are ridiculously hard to visualize because of its different variables and sizes. There are a few techniques discussed by [30] for dimensionality reduction because with large datasets, many dimensions can be seen and plotted to the interpretation could be complex and with full of errors. The high dimensional datasets have a lot of concerns for ML Algorithms such as increased computation time, storage space and pipeline performance. The techniques are Principal Component Analysis (PCA), ICA, Isometric Mapping (Isomap), LDA, UMAP, Locally Linear Embedding (LLE), And Autoencoders. These can be used to improve visualization and interpretation.
    a. *Principal Component Analysis:* This is used to reduce the number of features while still capturing the key information as measured by variance. It generates a new smaller feature set by combining the prominent features within the original set, called the principal components [31].
    b. *Independent Component Analysis (ICA):* This is a linear dimension reduction method, where the big dataset is converted to columns of independent components. ICA is also called as "Blind Source Separation" or "Cocktail Party Problem." [32]
    c. *Isometric Mapping (IsoMap):* This is a non-linear dimensionality reduction method. Its main motive is to reveal complex patterns in high dimension data by mapping them in lower-dimensional space while supporting relationships with all data points [33].
    d. *Linear Discriminant Analysis (LDA):* LDA, also known as Normal Discriminant Analysis or Discriminant Function Analysis. This is used to simplify the data by reducing the number of features or dimensions so that the distinct groups and classes are separated well from each other. [34]
    e. *Autoencoders:* This is composed of two parts an encoder and a decoder. Here, the encoder takes the input data and transforms it into a lower-dimensional representation, called the latent space or the bottleneck. The decoder takes the latent space and tries to reconstruct the original input data as closely as possible. The autoencoder learns to refine this process by minimizing the reconstruction error, or the difference between the input and the output [35].
2. *Challenge 2 Complexity:* Ali [37] states that the ability to understand patterns hidden in complex data is both a strength and a weakness of automated decision-making systems. An AI model does have many complex patterns that are hardly understood by data scientists or machine learning experts. Although AI algorithms are capable of extracting correlations across a wide range of complicated data, there is no assurance that these correlations are relevant or relate to real causal connections.

   The complexity could be managed by:

- These complications can be fixed by highlighting features influencing decisions making it more interpretable for the users simplifying the explanation process by focusing on key contributors.
- Creating a modular architecture that allows the separation of complex modules or components enabling the users/developer to focus more on which part of the model should be focused on.
- The study [38] focuses on Quantitative Evaluation to evaluate XAI complexity. Quantitative Evaluation/Metrics provide an objective way to minimize the influence of subjective humans' judgment. Common metrics include Depth of decision trees, Mean Reciprocal Rank, Runtime.

    (a) *Depth of decision trees:* These measure the complexity of the model; deeper trees indicate more intricate decision-making process.

    (b) *Mean Reciprocal Rank:* Features are ranked according to their significance; the lower the rank, better the explainability.

    (c) *Runtime:* The runtime evaluates XAI approaches' computing efficiency, which is important for real-world applications.

 doi:10.20944/preprints202410.0262.v1

3. *Challenge 3 Verification:* Evaluating the precision and thoroughness of XAI explanations is a complex challenge, particularly in the high stakes' domain like the healthcare as it deals with the patient's data [39]. Brandt et al. [40] developed synthetic classification models with the ground truth explanations [41] that serve as a reference for evaluating the accuracy for XAI models. These synthetic models significantly help in the verification within the XAI Models. They provide a controlled environment where the relationship between input or outputs are defined which allows the creation of ground truth (GT) explanations. This setup eliminates the biases associated with the real-world data, such as noise or unintentional correlations, making sure that the evaluations are made on clear criteria. The accuracy of the XAI methods can be mathematically represented with two metrics precision and recall [52].

   a. Precision (P): Precision measures how many explanations from the given list of explanations are correct. It is a way to quantify the correctness of explanations:

$$P = \frac{TP}{TP + FP}$$

- TP (True Positives): Correct Explanations that match the ground truth.
- FP (False Positives): Explanations that the AI model gave but do not match the ground truth.

The higher precision means that most of the answers/explanations given by XAI is correct indicating the XAI method is good and generates accurate explanations.

   b. Recall (R): Recall measures how many of the correct explanations (as per the ground truth) were identified by the XAI method. It indicates how well the method finds the explains all the relevant aspects of the model's decision.

$$P = \frac{TP}{TP + FN}$$

- FN (False Negatives): Correct explanations that were missed by XAI method.

High recall means that the XAI method identifies and explains most of the relevant aspects, even if it includes some incorrect explanations.

   c. F1 Score is the harmonic mean of Precision and Recall which provides single metric that balances both [53].

$$F1 = 2 \times \frac{P \times R}{P + R}$$

Consistency means that the important values of the features will stay the same even when the data is changed. Consistent explanations will help the users trust AI model by making sure that the answers are dependable and not arbitrary [42]. One such example for comparing the similarity and consistency between two given solutions is the Jaccard Similarity Index [54]. If an AI model provides explanations for similar instances or variations of an instance, the Jaccard Index can be used to compare the similarity of these explanations. A high Jaccard similarity would make sure that the explanations are consistent.

Jaccard Similarity = (number of observations in both sets) / (number in either set)

$$\mathbf{J\ (A, B) = |A \cap B| / |A \cup B|}$$

If two datasets have similarity, then their Jaccard Similarity Index will be 1 and if there are no common members then Jaccard Similarity index will be 0 [43].

Some things to make sure that AI models are consistent are:

1. Ensuring that the data is clean and free of errors. Poor data quality can lead to inconsistent model behaviour.
2. Normalization and standardization: Consistent preprocessing steps to all the data input is important. Normalization ensures that all the features contribute equally to the distance calculations like KNN and K-means.

 doi:10.20944/preprints202410.0262.v1

Sensitivity [55] presents significant challenges including variability in explanations, difficulty in assessing consistency, evaluating model behaviour, computational complexity, and overseeing adversarial perturbations. High sensitivity can lead to substantial changes in explanations with minor input variations, as quantified by:

$$\textbf{Sensitivity } (x, \delta) = \frac{1}{n}\Sigma_{i=1}^{n}|E(x) - E(x + \delta_i)|$$

where $E(x)$ is the explanation for input x and δ represents perturbations. This variability undermines the stability and reliability of the model's rationale, complicating the verification process. To assess consistency, one can use:

$$\textbf{Consistency } (x,x') = \frac{1}{m}\Sigma_{i=1}^{m}\left|E\left(x_j\right) - E\left(x + x_j'\right)\right|$$

where x_j and x'_j are inputs. This formula helps ensure explanations are consistent and reflective of meaningful patterns. Additionally, sensitivity can reveal model instability and reliance on noise rather than generalizable features, which can be assessed using:

$$\textbf{Feature Sensitivity} = \frac{1}{d}\Sigma_{k=1}^{d}|I(x_k) - I(x_k + \delta_k)|$$

where $I(x_k)$ is the importance of feature k and δ_k is the perturbation applied to it. The computational cost of conducting thorough sensitivity analyses adds to the complexity, which can be managed through efficient algorithms.

Adversarial perturbations [56] further expose vulnerabilities, impacting explanation robustness. Addressing these challenges involves implementing robust testing and consistency evaluation methods, using efficient algorithms for sensitivity analysis, and incorporating adversarial testing to ensure explanations remain stable and trustworthy.

Adversarial Sensitivity can be measured by how explanation change under adversarial attacks.

$$\textbf{Adversarial Sensitivity} = \frac{1}{p}\Sigma_{t=1}^{p} \mid E(x) - E(x + \delta_t) \mid$$

p is the number of adversarial perturbations and δ_t represents adversarial perturbations.

VI. Levels of Explanations

The study [49] explains about the five distinct levels of explanations for XAI, each representing a different aspect of how AI systems can communicate their reasoning and decision-making process. These levels are designed to align with human cognitive processes, ensuring that explanations are relatable and understandable to users, while addressing key areas such as contrastive explanation, attribution theory, and explanation selection. Structured hierarchically, the levels range from simple, reactive explanations (Zero-order) to more complex, reflective explanations (Meta) [fig 9], allowing for a bottom-up approach in generating explanations that can be refined iteratively based on user feedback and contextual understanding. The framework aims to develop AI systems that provide accurate, socially, and culturally relevant explanations, thereby enhancing user trust and engagement.

 doi:10.20944/preprints202410.0262.v1

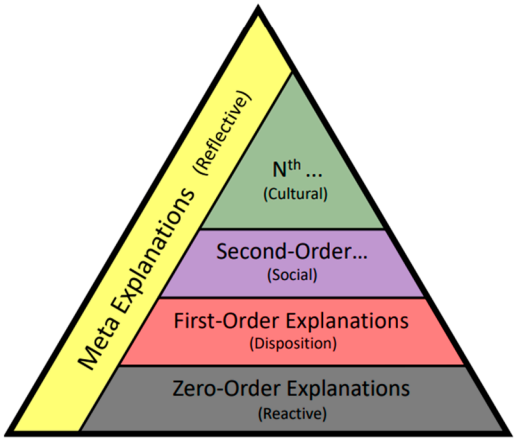

Figure 9. Levels of Explanation for XAI, providing a bottom-up constructivist model for explaining AI agent behaviour [49]. This model is adapted from Animal Cognitive Ethology's levels of intentionality [50,51].

A. Zero-order (Reactive)

Zero-order explanations are the most basic form of explanation, focusing on the immediate responses of an AI system to specific inputs. These explanations provide a direct interpretation of the agent's actions based on its perception of the environment, detailing how particular features influenced a decision or behaviour. For instance, in the context of a self-driving car, a zero-order explanation might clarify what the vehicle perceived just before a collision, such as identifying an obstacle or a pedestrian, thereby offering insight into the agent's reactive decision-making process.

B. First-order (Disposition)

First-order explanations delve deeper into the internal states and motivations of the AI agent, providing insights into the reasoning behind its decisions. This level addresses the agent's goals, beliefs, and intentions, explaining how these factors influence its behaviour. For example, in an autonomous vehicle, a first-order explanation might describe why the car chose to slow down or change lanes, highlighting its programmed objectives, risk assessments, and the underlying logic that guided its actions in each situation.

C. Second-order (Social)

Second-order explanations consider the social context in which the AI operates, examining how interactions with other agents and humans influence its behaviour. This level emphasizes the importance of social norms, relationships, and influences that shape the agent's decision-making processes. For instance, an AI system designed for customer service may adjust its responses based on the user's emotional state or feedback, providing a second-order explanation that reflects the social dynamics at play and how they impact the agent's interactions.

D. Nth order (Cultural)

Nth-order explanations address the broader cultural factors that shape the agent's decisions and behaviours, encompassing societal norms, values, and expectations. This level highlights how cultural context influences the interpretation of situations and the agent's responses. For example, an AI designed for global customer interactions may adapt its communication style based on cultural expectations of politeness and formality, providing an nth-order explanation that reflects the cultural nuances affecting its behaviour and decision-making.

E. Meta (Reflective)

Meta explanations provide insights into the explanation generation process itself, detailing how the AI agent selects or constructs its explanations. This level encompasses the rationale behind choosing specific types of explanations and the underlying reasoning processes involved. For instance, a meta-explanation might clarify why a particular explanation was deemed most relevant for a user query,

 doi:10.20944/preprints202410.0262.v1

including considerations of context, prior knowledge, and the processes used to generate the original decision, thereby enhancing transparency, and understanding of the AI's reasoning.

These levels of explanation refer to the different layers of understanding at which an AI systems behaviour can be explained, on the other hand even with these levels of explanations there are some models that are not easily interpreted or transparent known as the black box models. Some examples of the black box models include deep neural networks and ensemble methods. The black-box models can achieve high accuracy, but they do lack transparency which poses challenges in critical fields like healthcare and finance. Those five levels of explanations discussed by (...) can be narrowed down to High level, Mid-level, and low-level explanations.

The high-level characterization of black box models provides global descriptions or pictures of the behaviour of the model which can suggest global tendencies without explaining precisely how it works. While high-level perspectives offer more general information about model behaviour by approaching the black box from different angles or contexts, mid-level reasons help fill the gap between a general overview and concrete information about feature by offering a way to show how those features interact and affect or alter the predictions. Finally, low-level explanations are the most granular and precise since they rely on tools such as LIME or SHAP that break down the specific predictions made by the individual features, therefore providing a detailed explanation of how the model works.

VII. Conclusions

The increase in the use of Artificial Intelligence (AI) in present and emerging applications introduced an important need for transparency and for the ability to explain the results. Currently, there is a youthful branch of study termed Explainable AI or XAI that endeavours to enable human beings to understand the Black Box of AI. This results in that people believe in each other, work together, and act appropriately in business fields including healthcare, finance, and auto-driving cars.

In this paper, the notion of XAI was examined with the emphasis on its significance and implementation difficulties concerning explainability. The basic levels of operation were explained in terms of the kinds of explanations that can be given, ranging from purely reactive explanations (Zero-order) to high-level, reflective explanations of why explanation generation occurs (Meta). On the technical side, we also discussed the black-box models' multi-faced nature as well as the subtle differences between the explanations required for both views on the model, with or without hierarchy.

Visualization, feature importance analysis, and counterfactuals are a few of the XAI methodologies that have been studied; the body of research on these techniques is expanding quickly. With the further development of XAI, artificial intelligence (AI) will become more accountable and reliable, depending on human input to solve complex challenges. Future research may focus on issues such as XAI difficulties and directions depending on the design, development, and deployment phases of the machine learning life cycle. In addition to laying the foundation for future research on the connection between AI and enhanced trust in collaborative task contexts, this paper highlights the importance of suitable explanation strategies.

References

1. P. S. Kulkarni and A. S. N, "History and growth of artificial intelligence," *Indian Scientific Journal of Research in Engineering and Management*, vol. 07, no. 10, pp. 1–11, Oct. 2023, doi: 10.55041/ijsrem26432.
2. S. N. Srihari, "Explainable Artificial Intelligence: An overview," JSTOR, pp. 9–38, [Online]. Available: https://www.jstor.org/stable/27130153
3. Scott, Clancey, R. Davis, and Shortliffe, "Explanation Capabilities of production-based consultation systems," *American Journal of Computational Linguistics*, 1977.
4. R. Tiwari, "Explainable AI (XAI) and its Applications in Building Trust and Understanding in AI Decision Making," Indian Scientific Journal of Research in Engineering and Management, vol. 07, no. 01, Jan. 2023, doi: 10.55041/ijsrem17592.
5. I. Goodfellow, Y. Bengio, and A. Courville, Deep learning. MIT Press, 2016.
6. M. Wattenberg, F. Viégas, and I. Johnson, "How to Use t-SNE Effectively," vol. 1, no. 10, Oct. 2016, doi: 10.23915/distill.00002.
7. H. Vainio-Pekka et al., "The role of explainable AI in the research field of AI ethics," ACM Transactions on Interactive Intelligent Systems, vol. 13, no. 4, pp. 1–39, Dec. 2023, doi: 10.1145/3599974.

8. P. J. Phillips et al., "Four principles of explainable artificial intelligence," Sep. 2021. Doi: 10.6028/nist.ir.8312.
9. W. Saeed and C. Omlin, "Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities," Knowledge-based Systems, vol. 263, p. 110273, Mar. 2023, doi: 10.1016/j.knosys.2023.110273.
10. Wang, X. and Chen, X., 2024. Towards Human-AI Mutual Learning: A New Research Paradigm. arXiv preprint arXiv:2405.04687.
11. H. Felzmann, E. F. Villaronga, C. Lutz, and A. Tamò-Larrieux, "Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns," Big Data & Society, vol. 6, no. 1, p. 205395171986054, Jan. 2019, doi: 10.1177/2053951719860542.
12. T.-W. Chen and S. S. Sundar, "This App Would Like to Use Your Current Location to Better Serve You," Apr. 2018, doi: 10.1145/3173574.3174111.
13. M. Eslami, S. R. K. Kumaran, C. Sandvig, and K. Karahalios, "Communicating Algorithmic Process in Online Behavioral Advertising", Apr. 2018, doi: 10.1145/3173574.3174006.
14. E. Rader, K. Cotter, and J. Cho, "Explanations as Mechanisms for Supporting Algorithmic Transparency," Apr. 2018, doi: 10.1145/3173574.3173677.
15. R. Zhang et al., "I know this looks bad, but I can explain: understanding when AI should explain actions in Human-AI teams," ACM Transactions on Interactive Intelligent Systems, Dec. 2023, doi: 10.1145/3635474.
16. "Factorial Survey Design - Summer schools in Europe." https://www.summerschoolsineurope.eu/course/6467/factorial-survey-design
17. G. Ostuzzi, L. Benda, E. Costa, and C. Barbui, "The continuum of depressive experiences in cancer patients and the role of antidepressants. Results from a systematic review and meta-analysis," Journal of Psychosomatic Research, vol. 109, p. 124, Jun. 2018, doi: 10.1016/j.jpsychores.2018.03.114.
18. [18] S. Ali *et al.*, "Metaverse in Healthcare Integrated with Explainable AI and Blockchain: Enabling Immersiveness, Ensuring Trust, and Providing Patient Data Security," *Sensors*, vol. 23, no. 2, p. 565, Jan. 2023, doi: 10.3390/s23020565.
19. Z. Sadeghi *et al.*, "A brief review of explainable artificial intelligence in healthcare," *arXiv (Cornell University)*, Jan. 2023, doi: 10.48550/arxiv.2304.01543.
20. N. Rane, S. Choudhary, and J. Rane, "Explainable Artificial Intelligence (XAI) approaches for transparency and accountability in Financial Decision-Making," *Social Science Research Network*, Jan. 2023, doi: 10.2139/ssrn.4640316.
21. S. Nazat, L. Li, and M. Abdallah, "XAI-ADS: An Explainable artificial intelligence framework for enhancing anomaly detection in autonomous driving systems," *IEEE Access*, vol. 12, pp. 48583–48607, Jan. 2024, doi: 10.1109/access.2024.3383431.
22. S. Manna and N. Sett, "Need of AI in Modern Education: in the Eyes of Explainable AI (xAI)," *arXiv (Cornell University)*, Jul. 2024, doi: 10.48550/arxiv.2408.00025.
23. I. Abdulrashid, R. Z. Farahani, S. Mammadov, M. Khalafalla, and W.-C. Chiang, "Explainable artificial intelligence in transport Logistics: Risk analysis for road accidents," *Transportation Research Part E Logistics and Transportation Review*, vol. 186, p. 103563, Jun. 2024, doi: 10.1016/j.tre.2024.103563.
24. J. Černevičienė and A. Kabašinskas, "Explainable artificial intelligence (XAI) in finance: a systematic literature review," *Artificial Intelligence Review*, vol. 57, no. 8, Jul. 2024, doi: 10.1007/s10462-024-10854-8.
25. K. Demertzis, K. Rantos, L. Magafas, C. Skianis, and L. Iliadis, "A secure and Privacy-Preserving Blockchain-Based XAI-Justice system," *Information*, vol. 14, no. 9, p. 477, Aug. 2023, doi: 10.3390/info14090477.
26. E. Owens, B. Sheehan, M. Mullins, M. Cunneen, J. Ressel, and G. Castignani, "Explainable Artificial intelligence (XAI) in insurance," *Risks*, vol. 10, no. 12, p. 230, Dec. 2022, doi: 10.3390/risks10120230.
27. R. Branco *et al.*, "Explainable AI in Manufacturing: An Analysis of Transparency and Interpretability Methods for the XMANAI Platform," Jun. 2023, doi: 10.1109/ice/itmc58018.2023.10332373.
28. N. Nawaz, H. Arunachalam, B. K. Pathi, and V. Gajenderan, "The adoption of artificial intelligence in human resources management practices," *International Journal of Information Management Data Insights*, vol. 4, no. 1, p. 100208, Jan. 2024, doi: 10.1016/j.jjimei.2023.100208.
29. G. Makridis *et al.*, "XAI enhancing cyber defence against adversarial attacks in industrial applications," Dec. 2022, doi: 10.1109/ipas55744.2022.10052858.
30. R. Dwivedi *et al.*, "Explainable AI (XAI): core ideas, techniques, and solutions," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–33, Jan. 2023, doi: 10.1145/3561048.
31. K. Gross, "Dimensionality reduction (In plain English!)," Oct. 18, 2023. https://blog.dataiku.com/dimensionality-reduction-how-it-works-in-plain-english
32. W. Lu and J. C. Rajapakse, "ICA with Reference," *Neurocomputing*, vol. 69, no. 16–18, pp. 2244–2257, Oct. 2006, doi: 10.1016/j.neucom.2005.06.021.
33. B. Yang, M. Xiang, and Y. Zhang, "Multi-manifold Discriminant Isomap for visualization and classification," *Pattern Recognition*, vol. 55, pp. 215–230, Feb. 2016, doi: 10.1016/j.patcog.2016.02.001.

34. T. Xie, P. Qin, and L. Zhu, "Study on the topic Mining and Dynamic Visualization in view of LDA Model," *Modern Applied Science*, vol. 13, no. 1, p. 204, Dec. 2018, doi: 10.5539/mas.v13n1p204.
35. "What are the benefits and drawbacks of using autoencoders for dimensionality reduction?," *www.linkedin.com*, Jun. 23, 2024. https://www.linkedin.com/advice/0/what-benefits-drawbacks-using-autoencoders
36. S. M. Dafali, M. Kissi, and O. E. Beggar, "Comparative Study between Global and Local Explainable Models," Nov. 2023, doi: 10.1109/sita60746.2023.10373599.
37. S. Ali *et al.*, "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence," *Information Fusion*, vol. 99, p. 101805, Nov. 2023, doi: 10.1016/j.inffus.2023.101805.
38. N. A. M. Ahmed and A. Alpkoçak, "A quantitative evaluation of explainable AI methods using the depth of decision tree," *TURKISH JOURNAL OF ELECTRICAL ENGINEERING & COMPUTER SCIENCES*, vol. 30, no. 6, pp. 2054–2072, Sep. 2022, doi: 10.55730/1300-0632.3924.
39. Yang, Guang, Qinghao Ye and Jun Xia. "Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond." *An International Journal on Information Fusion* 77 (2021): 29 - 52.
40. R. Brandt, D. Raatjens, and G. Gaydadjiev, "Precise benchmarking of explainable AI attribution methods," *arXiv (Cornell University)*, Jan. 2023, doi: 10.48550/arxiv.2308.03161.
41. S. S. Amiri *et al.*, "Data representing Ground-Truth explanations to evaluate XAI methods," *arXiv (Cornell University)*, Jan. 2020, doi: 10.48550/arxiv.2011.09892.
42. M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': explaining the predictions of any classifier," *arXiv (Cornell University)*, Jan. 2016, doi: 10.48550/arxiv.1602.04938.
43. M. Jadeja, "Jaccard Similarity made simple: a beginner's guide to data comparison," *Medium*, Mar. 05, 2024. [Online]. Available: https://medium.com/@mayurdhvajsinhjadeja/jaccard-similarity-34e2c15fb524
44. B. Iglewicz and D. C. Hoaglin, *Volume 16: How to Detect and Handle Outliers*. Quality Press, 1993.
45. D. Pedreschi *et al.*, "Social AI and the challenges of the Human-AI ecosystem," *arXiv (Cornell University)*, Jan. 2023, doi: 10.48550/arxiv.2306.13723.
46. S. Reeves, J. Peller, J. Goldman, and S. Kitto, "Ethnography in qualitative educational research: AMEE Guide No. 80," *Medical Teacher*, vol. 35, no. 8, pp. e1365–e1379, Jun. 2013, doi: 10.3109/0142159x.2013.804977.
47. R. Budiu, "Between-Subjects vs. Within-Subjects Study Design," *Nielsen Norman Group*, Jan. 16, 2024. https://www.nngroup.com/articles/between-within-subjects/
48. A. Goyal, "Explainability Frameworks (XAI) Review on Structured data," *Medium*, Oct. 17, 2022. [Online].
49. R. Dazeley, P. Vamplew, C. Foale, C. Young, S. Aryal, and F. Cruz, "Levels of explainable artificial intelligence for human-aligned conversational explanations," *Artificial Intelligence*, vol. 299, p. 103525, May 2021, doi: 10.1016/j.artint.2021.103525.
50. D. R. Griffin, The Question Of Animal Awareness: Evolutionary Continuity Of Mental Experience, Rockefeller University Press, San Mateo, CA, 1976.
51. D. L. Cheney, R. M. Seyfarth, How Monkeys See The World: Inside the mind of another species, University of Chicago Press, Chicago and London, 1990.
52. S. S. Amiri *et al.*, "Data representing Ground-Truth explanations to evaluate XAI methods," *arXiv (Cornell University)*, Jan. 2020, doi: 10.48550/arxiv.2011.09892.
53. H. Huang, H. Xu, X. Wang, and W. Silamu, "Maximum F1-Score Discriminative training criterion for automatic mispronunciation detection," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 23, no. 4, pp. 787–797, Apr. 2015, doi: 10.1109/taslp.2015.2409733.
54. G. I. Ivchenko and S. A. Honov, "On the jaccard similarity test," *Journal of Mathematical Sciences*, vol. 88, no. 6, pp. 789–794, Mar. 1998, doi: 10.1007/bf02365362.
55. B. Van Stein, E. Raponi, Z. Sadeghi, N. Bouman, R. C. H. J. Van Ham, and T. Bäck, "A comparison of global sensitivity analysis methods for explainable AI with an application in genomic prediction," *IEEE Access*, vol. 10, pp. 103364–103381, Jan. 2022, doi: 10.1109/access.2022.3210175.
56. A. Hartl, M. Bachl, J. Fabini, and T. Zseby, "Explainability and Adversarial Robustness for RNNs," ., Aug. 2020, doi: 10.1109/bigdataservice49289.2020.00030.
57. I. X, "Use cases of explainable AI (XAI) across various sectors," *Medium*, Nov. 19, 2023. [Online]. Available: https://medium.com/@inspirexnewsletter/use-cases-of-explainable-ai-xai-across-various-sectors-ffa7d7fa1778