What’s beyond AI Explainability?

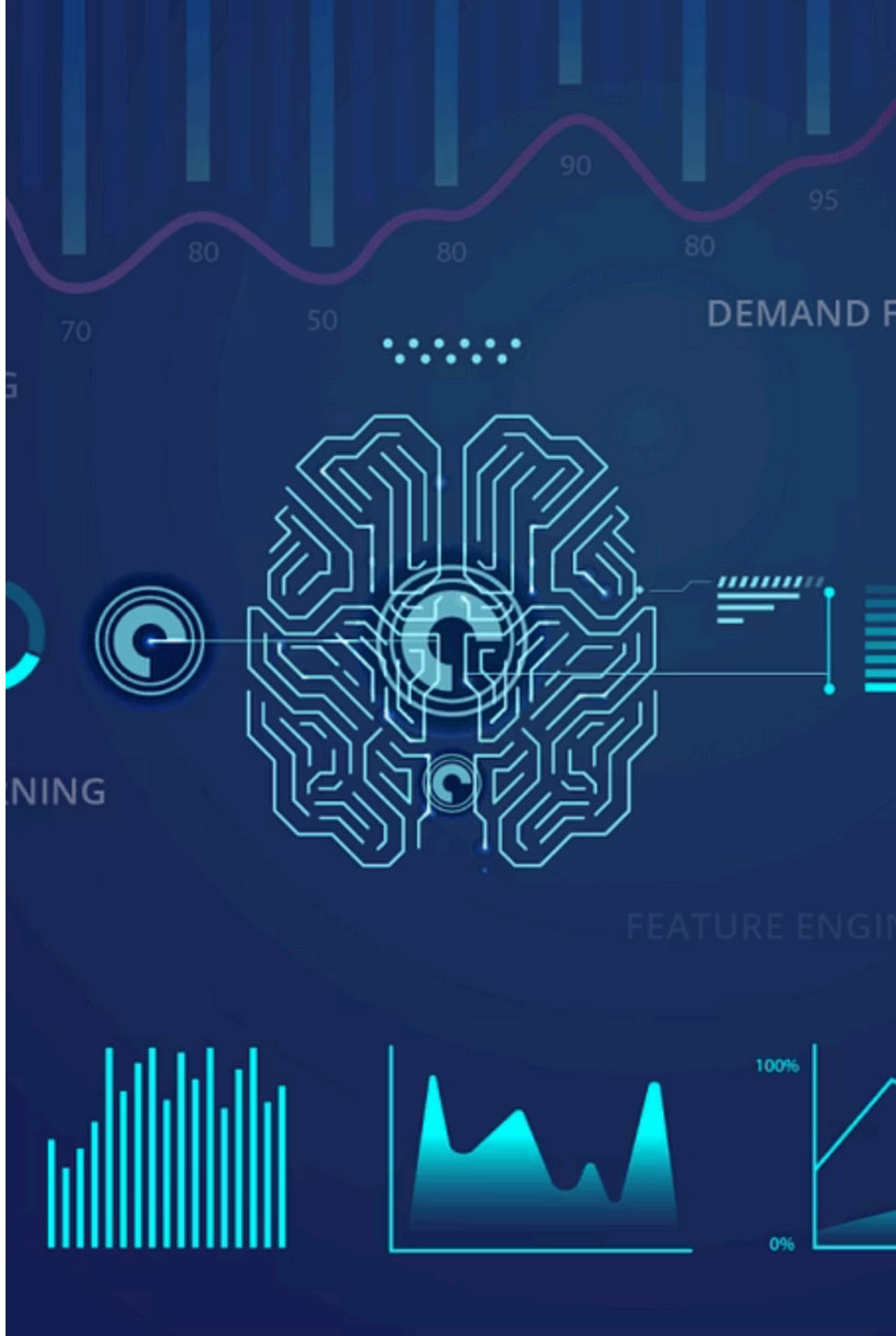
Explainable Artificial Intelligence

Presented By :

Ahmed Layouni

Overview

- Introduction to XAI
- Importance of XAI
- Comparing XAI and AI
- Benefits
- Techniques Used in XAI
- Challenges and Considirations
- Real-World Applications
- Limitations



What is XAI?

Definition: Explainable artificial intelligence (XAI) is a set of processes and methods that allows human users to comprehend and trust the results and output created by machine learning algorithms.

Importance: It aims to reveal the "why" behind AI's predictions, rather than just presenting the results. This involves using various techniques to make complex AI models more transparent and interpretable.

Why explainable AI matters ?

- It is crucial for an organization to have a full understanding of the AI decision-making processes with model monitoring and accountability of AI and not to trust them blindly.
- Explainable AI can help humans understand and explain machine learning (ML) algorithms, deep learning and neural networks.

Why explainable AI matters ?

- Explainable AI is one of the key requirements for implementing responsible AI, a methodology for the large-scale implementation of AI methods in real organizations with fairness, model explainability and accountability.

XAI vs AI

What exactly is the difference between “regular” AI and explainable AI?

XAI implements specific techniques and methods to ensure that each decision made during the ML process can be traced and explained. AI, on the other hand, often arrives at a result using an ML algorithm, but the architects of the AI systems do not fully understand how the algorithm reached that result.

This makes it hard to check for accuracy and leads to loss of control, accountability and auditability.

XAI vs AI

How does explainable AI relate to responsible AI?

First,Responsible AI refers to the development, deployment, and use of AI systems in a way that aligns with ethical principles, societal values, and legal standards.

Explainable AI looks at AI results after the results are computed.

Responsible AI looks at AI during the planning stages to make the AI algorithm responsible before the results are computed.

Explainable and responsible AI can work together to make better AI.

XAI vs AI

Explainability VS interpretability in AI:

Interpretability is the degree to which an observer can understand the cause of a decision. It is the success rate that humans can predict for the result of an AI output, while explainability goes a step further and looks at how the AI arrived at the result.

What are the benefits of Explainable AI?

- Build trust in production AI.
- Rapidly bring your AI models to production.
- Ensure interpretability and explainability of AI models.
- Simplify the process of model evaluation while increasing model transparency and traceability.

What are the benefits of Explainable AI?

- Systematically monitor and manage models to optimize business outcomes.
- Continually evaluate and improve model performance.
- Fine-tune model development efforts based on continuous evaluation.

What are the benefits of Explainable AI?

- Keep your AI models explainable and transparent.
- Manage regulatory, compliance, risk and other requirements
- Minimize overhead of manual inspection and costly errors.
- Mitigate risk of unintended bias.

Techniques Used in XAI

1. **Model-Agnostic Tools:**
   - **Local Interpretable Model-agnostic Explanations (LIME):**LIME provides explanations for individual predictions by approximating the complex model with a simpler, interpretable model in the vicinity of that prediction.
   - **SHapley Additive exPlanations (SHAP)**:SHAP assigns a value to each feature for a particular prediction, indicating its contribution to the prediction.

Techniques Used in XAI

2. Visualization: Visualizing the decision process, such as highlighting important regions in an image or showing how different inputs contribute to a prediction, can make the AI's logic more intuitive.

Tools:

- **Partial Dependence Plots (PDP)**
  - Show how one or two features affect predictions on average.
- **Individual Conditional Expectation (ICE) Plots**
  - Similar to PDP, but for individual data points.
- **Feature Importance Charts**
  - Rank features by their impact on the model.
- **Decision Trees**
  - Can be used alone or to approximate black-box models.

Techniques Used in XAI

3. Model-Specific Tools:

- **TreeExplainer**
  - A version of SHAP optimized for tree-based models (e.g., XGBoost, LightGBM).
- **Integrated Gradients (for neural networks)**
  - Attributes the prediction by integrating gradients of the model's output w.r.t. the input.
  - Often used in image and text models.
- **Grad-CAM / Saliency Maps (for CNNs)**
  - Visual tools that highlight image areas influencing classification decisions.

Techniques Used in XAI

4.Toolkits & Libraries:

- **IBM AI Explainability 360**
  - An open-source toolkit with various explainability metrics and methods.
- **Google's What-If Tool**
  - Provides visual interfaces to probe models, test counterfactuals, and analyze fairness.
- **Microsoft InterpretML**
  - Combines glass-box models (like GAMs) and black-box explanation tools (like SHAP, LIME).
- **Captum (by PyTorch)**
  - Library for interpretability in deep learning models using PyTorch.
- **Eli5**
  - Debugging and explanation for scikit-learn, XGBoost, and others.

Techniques Used in XAI

- **Traceability:**An example of a traceability XAI technique is DeepLIFT (Deep Learning Important FeaTures), which compares the activation of each neuron to its reference neuron and shows a traceable link between each activated neuron and even shows dependencies between them.

Challenges & Considerations

- **Fairness and debiasing**: Manage and monitor fairness. Scan the deployment for potential biases.
- **Model drift mitigation**: Analyze your model and make recommendations based on the most logical outcome. Alert when models deviate from the intended outcomes.
- **Model risk management**: Quantify and mitigate model risk. Get alerted when a model performs inadequately. Understand what happened when deviations persist.

Challenges & Considerations

- **Lifecycle automation:** Build, run and manage models as part of integrated data and AI services. Unify the tools and processes on a platform to monitor models and share outcomes. Explain the dependencies of machine learning models.

- **Multicloud-readiness:** Deploy AI projects across hybrid clouds including public clouds, private clouds and on premises. Promote trust and confidence with explainable AI.

Real-World Applications

- **Healthcare:** Accelerate diagnostics, image analysis, resource optimization and medical diagnosis. Improve transparency and traceability in decision-making for patient care. Streamline the pharmaceutical approval process with explainable AI.

- **Financial services:** Improve customer experiences with a transparent loan and credit approval process. Speed credit risk, wealth management and financial crime risk assessments. Accelerate resolution of potential complaints and issues. Increase confidence in pricing, product recommendations and investment services.

Real-World Applications

- **Autonomous Vehicles:** Help understand why a self-driving car made a specific maneuver, which is critical for safety and trust.

- **Fraud Detection:** Help identify the factors that contribute to fraudulent transactions, allowing for better fraud prevention.

Limitations of XAI

Explainable AI faces several limitations, including trade-offs between accuracy and interpretability, the potential for biased explanations, and challenges in scaling and computational complexity. Furthermore, XAI methods may not always capture the full complexity of decision-making processes, especially in dynamic or context-dependent situations.