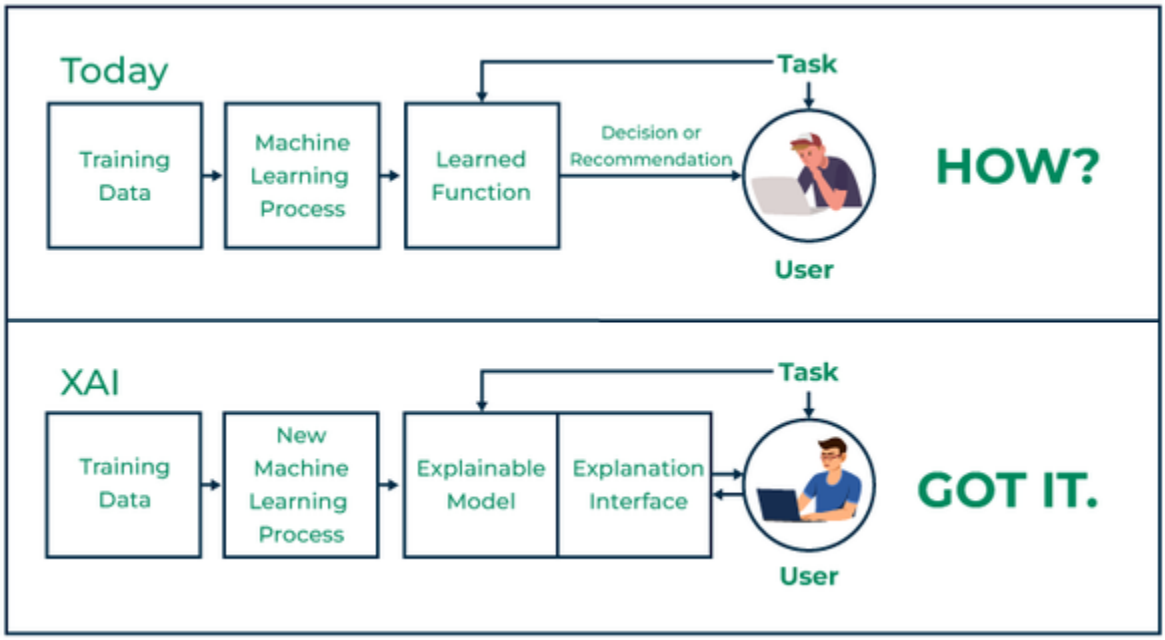
Explainable artificial intelligence (XAI) refers to a collection of procedures and techniques that enable machine learning algorithms to produce output and results that are understandable and reliable for human users. Explainable AI is a key component of the fairness, accountability, and transparency (FAT) machine learning paradigm and is frequently discussed in connection with deep learning. Organizations looking to establish trust when deploying AI can benefit from XAI. XAI can assist them in comprehending the behavior of an AI model and identifying possible problems like AI.



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.

ML models are often thought of as black boxes that are impossible to interpret.[2] Neural networks used in deep learning are some of the hardest for a human to understand. Bias, often based on race, gender, age or location, has been a long-standing risk in training AI models. Further, AI model performance can drift or degrade because production data differs from training data. This makes it crucial for a business to

continuously monitor and manage models to promote AI explainability while measuring the business impact of using such algorithms. Explainable AI also helps promote end user trust, model auditability and productive use of AI. It also mitigates compliance, legal, security and reputational risks of production AI.

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.[3] To help adopt AI responsibly, organizations need to embed ethical principles into AI applications and processes by building AI systems based on trust and transparency.

Origin of Explainable AI

The origins of explainable AI can be traced back to the early days of machine learning research when scientists and engineers began to develop algorithms and techniques that could learn from data and make predictions and inferences. As machine learning algorithms became more complex and sophisticated, the need for transparency and interpretability in these models became increasingly important, and this need led to the development of explainable AI approaches and methods.

One of the key early developments in explainable AI was the work of **Judea Pearl,** who introduced the concept of **causality in machine learning**, and proposed a framework for understanding and explaining the factors that are most relevant and influential in the model's predictions. This work laid the foundation for many of the explainable AI approaches and methods that are used today and provided a framework for transparent and interpretable machine learning.

Another important development in explainable AI was the work of **LIME (Local Interpretable Model-agnostic Explanations)**, which introduced a

method for providing interpretable and explainable machine learning models. This method uses a local approximation of the model to provide insights into the factors that are most relevant and influential in the model's predictions and has been widely used in a range of applications and domains.

Benefits of explainable AI

The value of explainable AI lies in its ability to provide transparent and interpretable machine-learning models that can be understood and trusted by humans. This value can be realized in different domains and applications and can provide a range of benefits and advantages. Some of the key values of explainable AI include:

1. Improved decision-making:- Explainable AI can provide valuable insights and information that can be used to support and improve decision-making. For example, explainable AI can provide insights into the factors that are most relevant and influential in the model's predictions, and can help to identify and prioritize the actions and strategies that are most likely to achieve the desired outcome.

2. Increased trust and acceptance:- Explainable AI can help to build trust and acceptance of machine learning models, and can overcome the challenges and limitations of traditional machine learning models, which are often opaque and inscrutable. This increased trust and acceptance can help to accelerate the adoption and deployment of machine learning models and can provide valuable insights and benefits in different domains and applications.

3. Reduced risks and liabilities:- Explainable AI can help to reduce the risks and liabilities of machine learning models, and can provide a framework for addressing the regulatory and ethical considerations of this technology. This reduced risk and liability can help to mitigate the potential impacts and

consequences of machine learning, and can provide valuable insights and benefits in different domains and applications.

Overall, the value of explainable AI lies in its ability to provide transparent and interpretable machine-learning models that can be understood and trusted by humans. This value can be realized in different domains and applications and can provide a range of benefits and advantages.

How explainable AI works

With explainable AI as well as interpretable machine learning, organizations can gain access to AI technology's underlying decision-making and are empowered to make adjustments. Explainable AI can improve the user experience of a product or service by helping the end user trust that the AI is making good decisions. When do AI systems give enough confidence in the decision that you can trust it, and how can the AI system correct errors that arise?[4]

As AI becomes more advanced, ML processes still need to be understood and controlled to ensure AI model results are accurate. Let's look at the difference between AI and XAI, the methods and techniques used to turn AI to XAI, and the difference between interpreting and explaining AI processes.

Comparing AI and XAI

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.

Explainable AI techniques

The setup of XAI techniques consists of three main methods. Prediction accuracy and traceability address technology requirements while decision understanding addresses human needs. Explainable AI especially explainable machine learning, will be essential

if future warfighters are to understand, appropriately trust, and effectively manage an emerging generation of artificially intelligent machine partners.[5]

Prediction accuracy

Accuracy is a key component of how successful the use of AI is in everyday operation. By running simulations and comparing XAI output to the results in the training data set, the prediction accuracy can be determined. The most popular technique used for this is Local Interpretable Model-Agnostic Explanations (LIME), which explains the prediction of classifiers by the ML algorithm.

Traceability

Traceability is another key technique for accomplishing XAI. This is achieved, for example, by limiting the way decisions can be made and setting up a narrower scope for ML rules and features. 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.

Decision understanding

This is the human factor. Many people have a distrust in AI, yet to work with it efficiently, they need to learn to trust it. This is accomplished by educating the team working with the AI so they can understand how and why the AI makes decisions.

Five considerations for explainable AI

To drive desirable outcomes with explainable AI, consider the following.

Fairness and debiasing: Manage and monitor fairness. Scan your 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.

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.

Use cases for explainable AI

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.

Criminal justice: Optimize processes for prediction and risk assessment. Accelerate resolutions using explainable AI on DNA analysis, prison population analysis and crime forecasting. Detect potential biases in training data and algorithms.