

Artificial Intelligence (AI)

Machine Learning (ML)

Neural Networks (NNs)

Deep Learning (DL)

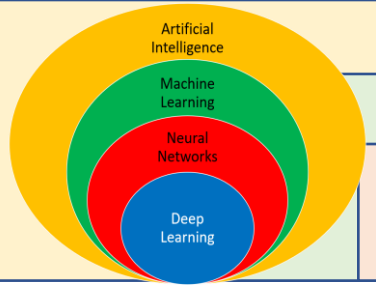
Advanced Deep Learning

Dr. Rastgoo









Artificial Intelligence (AI)

Machine Learning (ML)

Neural Networks (NNs)

Deep Learning (DL)

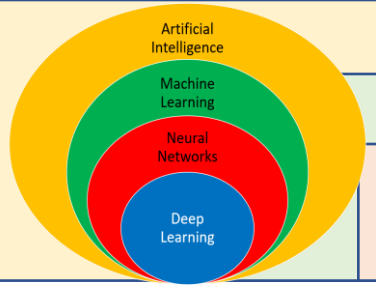
Introduction to Explainable Artificial Intelligence (XAI)

- Artificial Intelligence (AI) systems, particularly those powered by deep learning, have achieved remarkable success across domains such as computer vision, natural language processing, healthcare, finance, and autonomous systems.

- Despite these achievements, the decision-making processes of many modern AI models remain opaque, often described as "**black boxes**", due to their complex architectures and high-dimensional feature representations.

- This opacity has created a critical barrier to trust, accountability, and adoption in safety-critical and socially sensitive applications.

- To address this challenge, the field of Explainable Artificial Intelligence (XAI) has emerged as a research paradigm aimed at making AI systems transparent, interpretable, and trustworthy without significantly compromising predictive performance.



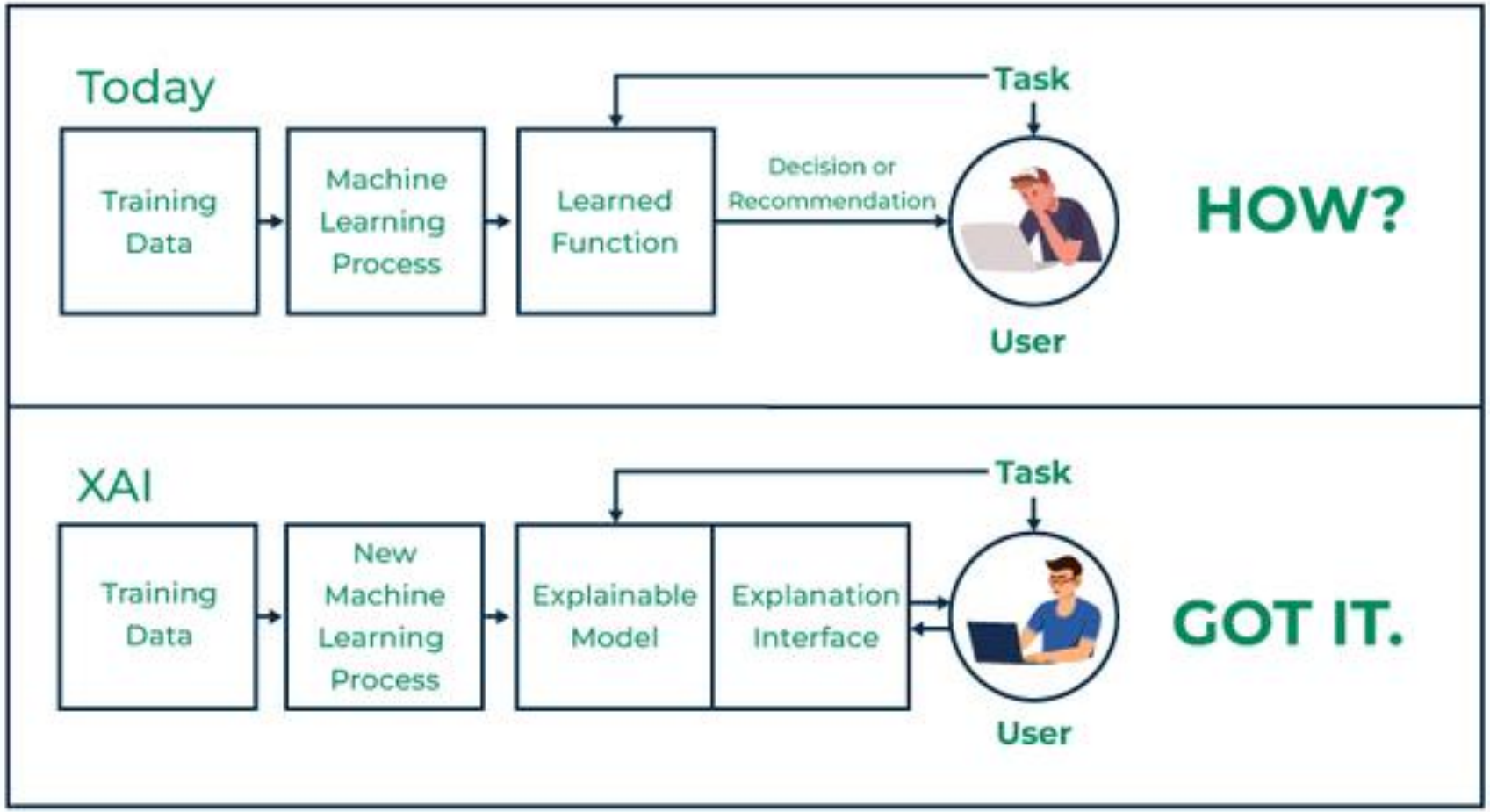
Artificial Intelligence (AI)

Machine Learning (ML)

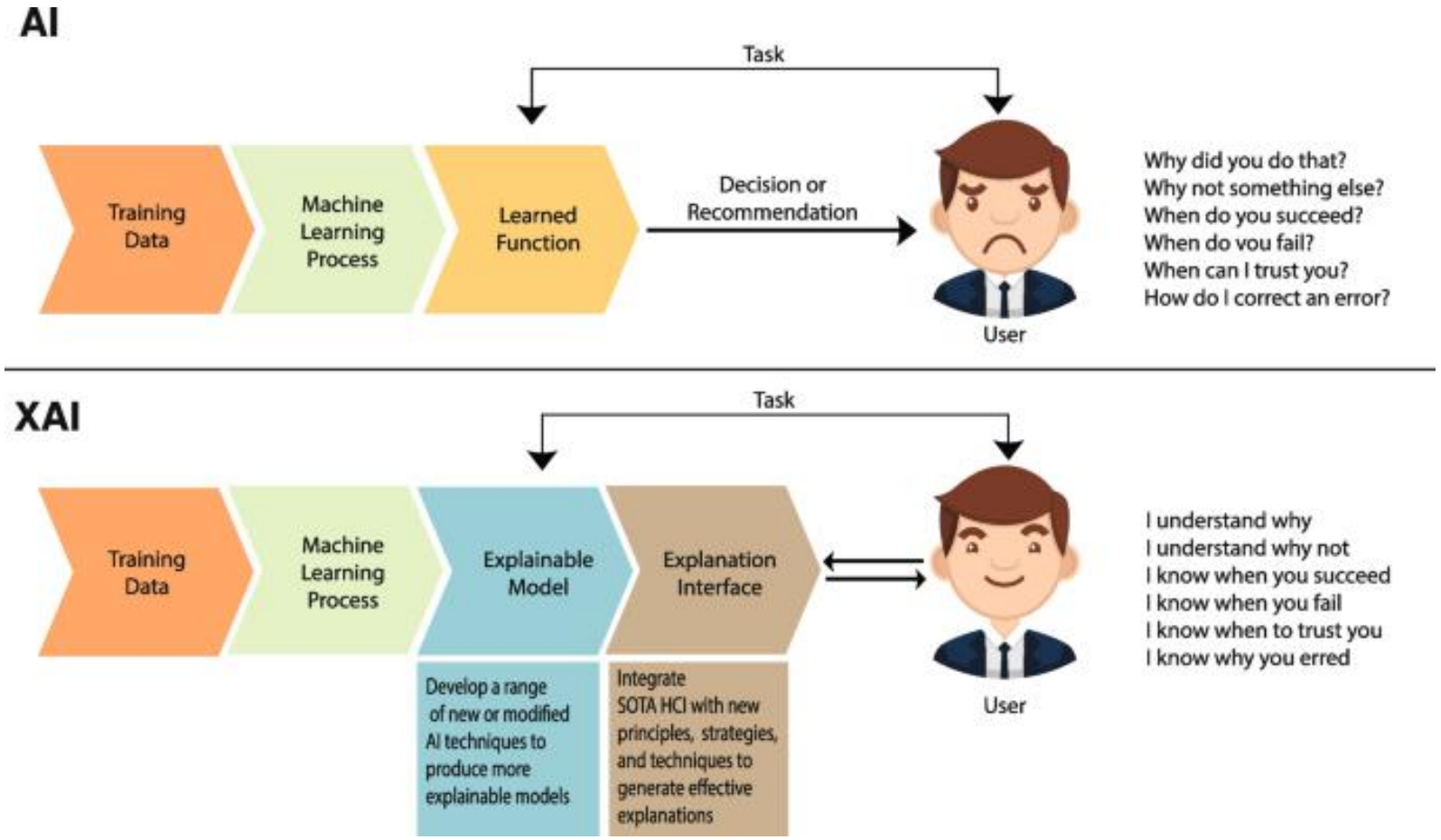
Neural Networks (NNs)

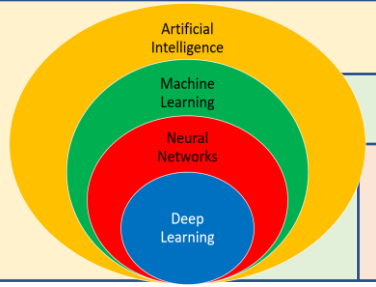
Deep Learning (DL)

Introduction to Explainable Artificial Intelligence (XAI)



Introduction to Explainable Artificial Intelligence (XAI)





Artificial Intelligence (AI)

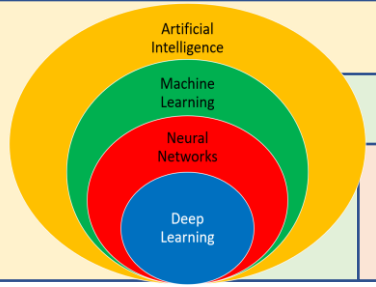
Machine Learning (ML)

Neural Networks (NNs)

Deep Learning (DL)

Motivation and Need for Explainability

- The need for explainability arises from both technical and ethical imperatives.
- Technically, XAI enables researchers and practitioners to understand model behavior, detect errors, detect biases, and improve model robustness.
- Ethically and legally, explainability supports human oversight, fairness, and compliance with regulatory frameworks such as the European Union's General Data Protection Regulation (GDPR), which grants individuals the "right to explanation" for automated decisions.
- In domains like healthcare and autonomous driving, where AI decisions may have life-altering consequences, interpretability is not optional but essential.
- Furthermore, explainability fosters human-AI collaboration, where users can trust and effectively interact with intelligent systems by understanding their rationale and limitations.



Artificial Intelligence (AI)

Machine Learning (ML)

Neural Networks (NNs)

Deep Learning (DL)

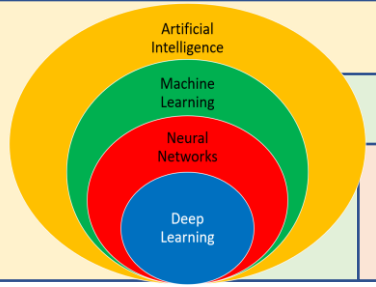
Interpretability vs. Explainability

- While the terms interpretability and explainability are often used interchangeably, they embody nuanced differences.

- Interpretability refers to the degree to which a human can understand the internal mechanics of a model directly; for example, the coefficients in a linear regression model or decision paths in a decision tree.

- Explainability, in contrast, extends beyond mere transparency: it encompasses the ability to provide meaningful and human-understandable explanations for model outputs, often through post-hoc analysis of complex models like deep neural networks.

- Thus, interpretability is often associated with inherently transparent models, whereas explainability involves methods that approximate or reconstruct human-understandable reasoning around opaque systems.



Artificial Intelligence (AI)

Machine Learning (ML)

Neural Networks (NNs)

Deep Learning (DL)

The Taxonomy of XAI Approaches

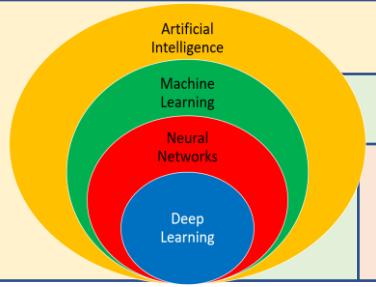
❖ Explainable AI methods can broadly be categorized based on when and how the explanations are generated:

❖ Intrinsic vs. Post-Hoc Explainability:

✓ Intrinsic (Model-Based) Explainability: Models are designed to be interpretable by nature. Examples include linear regression, decision trees, and attention-based architectures. These models trade off complexity for interpretability.

✓ Post-Hoc Explainability: These methods analyze already-trained black-box models to produce human-understandable explanations. Techniques include saliency maps, feature importance, surrogate models, and counterfactual reasoning.

Surrogate models: How can we approximate a complex AI model with a simpler, more interpretable model?
Counterfactual reasoning: What would need to change in the input for the AI model to produce a different prediction or decision?



Artificial Intelligence (AI)

Machine Learning (ML)

Neural Networks (NNs)

Deep Learning (DL)

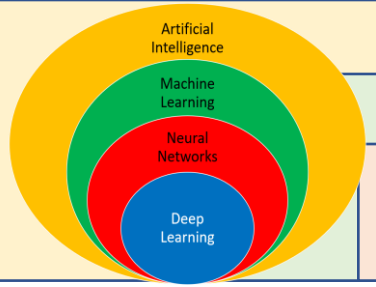
The Taxonomy of XAI Approaches

❖ Explainable AI methods can broadly be categorized based on when and how the explanations are generated:

❖ Global vs. Local Explanations

✓ Global Explainability: Aims to describe the model's overall behavior and decision logic. Example: feature importance across an entire dataset.

✓ Local Explainability: Focuses on explaining individual predictions or instances. Example: local linear approximations or perturbation-based explanations (e.g., LIME, SHAP).



Artificial Intelligence (AI)

Machine Learning (ML)

Neural Networks (NNs)

Deep Learning (DL)

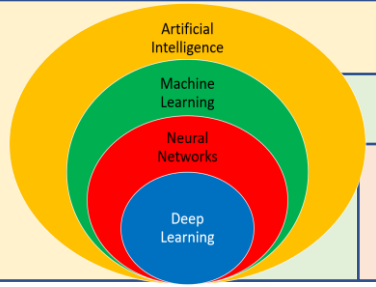
The Taxonomy of XAI Approaches

❖ Explainable AI methods can broadly be categorized based on when and how the explanations are generated:

❖ Model-Agnostic vs. Model-Specific Approaches

✓ Model-Agnostic Methods: Can be applied to any machine learning model, regardless of its internal architecture. Examples include LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations).

✓ Model-Specific Methods: Tailored to particular architectures. For instance, Grad-CAM (Gradient-weighted Class Activation Mapping) is designed specifically for convolutional neural networks (CNNs), while Integrated Gradients and Attention Visualization are used for transformer-based models.



Artificial Intelligence (AI)

Machine Learning (ML)

Neural Networks (NNs)

Deep Learning (DL)

The Taxonomy of XAI Approaches

❖ Explainable AI methods can broadly be categorized based on when and how the explanations are generated:

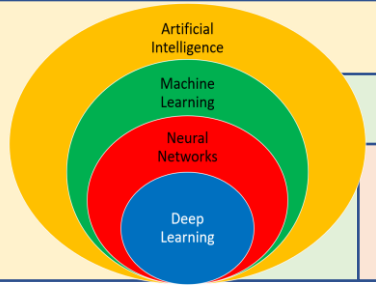
Shapley values are a concept from cooperative game theory used to measure how much each participant contributes to the outcome of a cooperative group. In explainable artificial intelligence (XAI), they are used to quantify how much each feature contributes to a model's prediction.

❖ Prominent XAI Techniques

1. Feature Attribution Methods

Feature attribution quantifies how much each input feature contributes to a model's prediction. Examples:

- ✓ SHAP: Based on cooperative game theory, it assigns Shapley values to features reflecting their contribution to the outcome.
- ✓ LIME: Fits a simple surrogate model locally around a prediction to approximate its decision boundary.
- ✓ Integrated Gradients: Computes the integral of gradients of the output with respect to the input along a path from a baseline to the actual input.



Artificial Intelligence (AI)

Machine Learning (ML)

Neural Networks (NNs)

Deep Learning (DL)

The Taxonomy of XAI Approaches

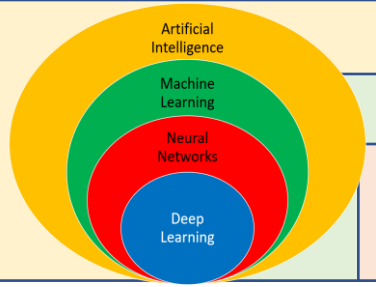
❖ Explainable AI methods can broadly be categorized based on when and how the explanations are generated:

❖ Prominent XAI Techniques

2. Visualization-Based Explanations

These methods provide visual insights into what parts of the input influence the model's output:

- ✓ Saliency Maps: Highlight regions of an image most relevant to a classification.
- ✓ Grad-CAM: Produces heatmaps to visualize spatial importance in CNNs.
- ✓ Attention Maps: Used in Transformer architectures to visualize interdependencies among tokens.



Artificial Intelligence (AI)

Machine Learning (ML)

Neural Networks (NNs)

Deep Learning (DL)

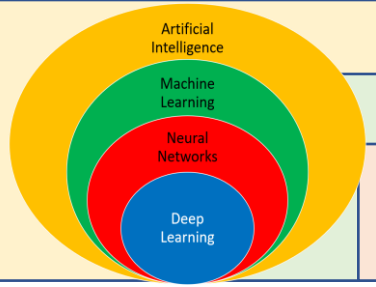
The Taxonomy of XAI Approaches

❖ Explainable AI methods can broadly be categorized based on when and how the explanations are generated:

❖ Prominent XAI Techniques

3. Example-Based and Counterfactual Explanations

- ✓ Prototype and Criticism Methods: Identify representative and atypical examples from the dataset.
- ✓ Counterfactual Explanations: Provide alternative input scenarios that would have changed the model's decision, useful for decision accountability and fairness.



Artificial Intelligence (AI)

Machine Learning (ML)

Neural Networks (NNs)

Deep Learning (DL)

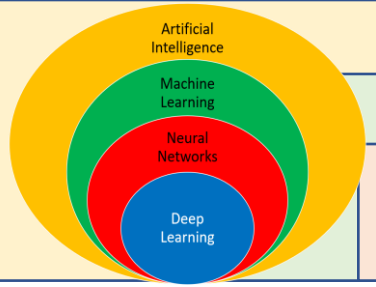
The Taxonomy of XAI Approaches

❖ Explainable AI methods can broadly be categorized based on when and how the explanations are generated:

❖ Prominent XAI Techniques

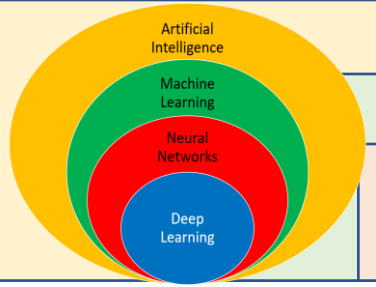
4. Concept-Based Explanations

- ✓ TCAV (Testing with Concept Activation Vectors): Measures the sensitivity of predictions to high-level human concepts rather than raw features.
- ✓ Disentangled Representation Learning: Encourages the model to encode interpretable factors of variation in the latent space.



Evaluation of Explainability

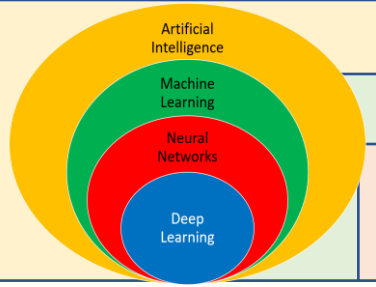
- Quantifying explainability remains a significant challenge.
- Current evaluation frameworks balance fidelity (the accuracy of the explanation in representing the model's behavior) and interpretability (how understandable it is to humans).
- Other metrics include completeness, stability, and usefulness in decision-making contexts.
- Human-centered evaluation, through user studies and domain expert assessments, is increasingly employed to complement quantitative measures.



Challenges and Limitations

❖ Despite its progress, XAI faces multiple challenges:

- ✓ Trade-off Between Accuracy and Interpretability: Highly interpretable models may underperform compared to deep learning models.
- ✓ Scalability: Many explanation methods struggle with large-scale or high-dimensional data.
- ✓ Subjectivity and Human Bias: Interpretations can vary depending on the user's expertise or expectations.
- ✓ Causality vs. Correlation: Most XAI techniques are correlation-based and fail to capture causal reasoning.
- ✓ Adversarial Manipulation: Explanations themselves can be manipulated or optimized to mislead users (the "explanation game").



Artificial Intelligence (AI)

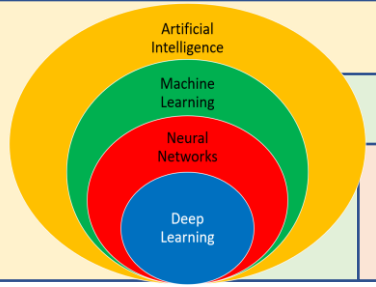
Machine Learning (ML)

Neural Networks (NNs)

Deep Learning (DL)

Emerging Directions in XAI

- Recent trends in XAI research focus on:
  - ✓ Causality-Driven Explanations: Integrating causal inference for more robust and reliable interpretability.
  - ✓ Neuro-Symbolic Explainability: Combining neural networks with symbolic reasoning for human-comprehensible logic.
  - ✓ Explainability in Generative AI: Understanding and controlling outputs of large-scale generative models (e.g., diffusion models, LLMs).
  - ✓ Human-Centered XAI: Designing explanations tailored to the user's context, expertise, and cognitive preferences.
  - ✓ Faithful and Robust Explanation Models: Ensuring that explanations reflect true model behavior and remain consistent under perturbations.



Artificial Intelligence (AI)

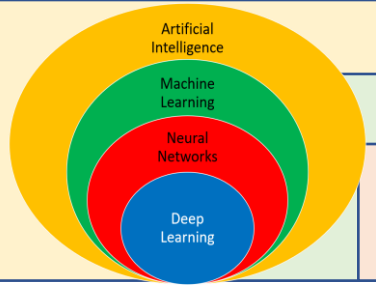
Machine Learning (ML)

Neural Networks (NNs)

Deep Learning (DL)

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



Artificial Intelligence (AI)

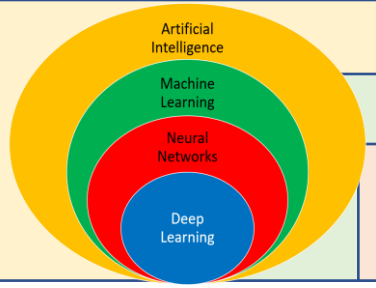
Machine Learning (ML)

Neural Networks (NNs)

Deep Learning (DL)

Conclusion

- Explainable Artificial Intelligence stands at the intersection of machine learning, cognitive science, human-computer interaction, and ethics.
- It aims not only to unveil the inner workings of AI systems but also to foster trust, accountability, and transparency in their deployment.
- As AI continues to permeate critical societal functions, XAI will play a pivotal role in bridging the gap between algorithmic intelligence and human understanding.
- The future of trustworthy AI depends on achieving a symbiotic balance between performance and interpretability, ensuring that AI systems remain both powerful and comprehensible.



Artificial Intelligence (AI)

Machine Learning (ML)

Neural Networks (NNs)

Deep Learning (DL)

