Explainable AI and Explaining AI

Prisha Baveja
Radhika Garg
Ritika Thakur

May 2024

Contents

Introduction To Explainable AI

- Helps us to know the 'why' rather than just the 'what'

Introduction To Explainable AI

- Helps us to know the 'why' rather than just the 'what'
- Refers to the degree to which humans can comprehend a model's result

Introduction To Explainable AI

- Helps us to know the 'why' rather than just the 'what'
- Refers to the degree to which humans can comprehend a model's result
- Difference between Explainability and Interpretability
- Decision tree is a typical method designed with explainable structure.

Introduction To Explainable AI

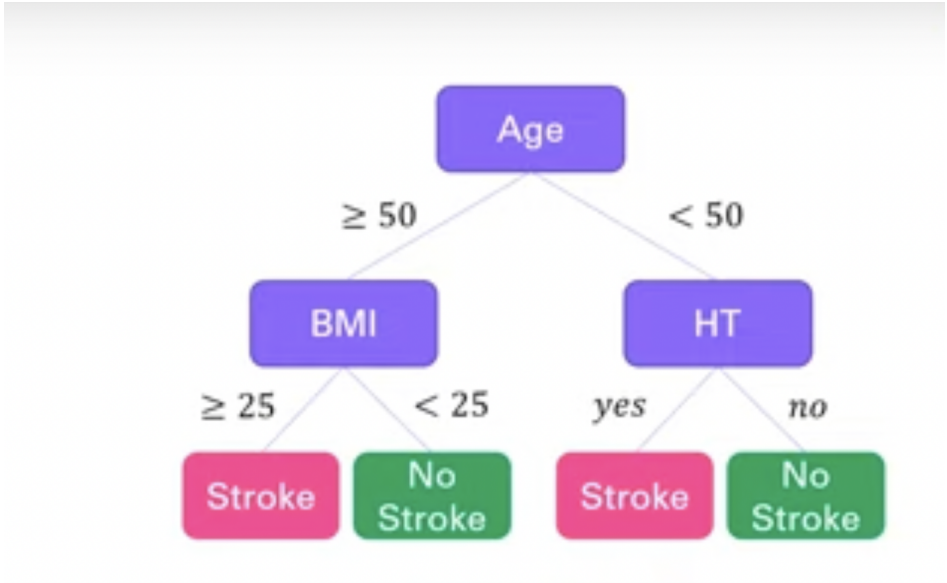

Figure: An example of decision tree, used by starting at the top and going down, level by level, according to the defined logic.

Introduction To Explainable AI

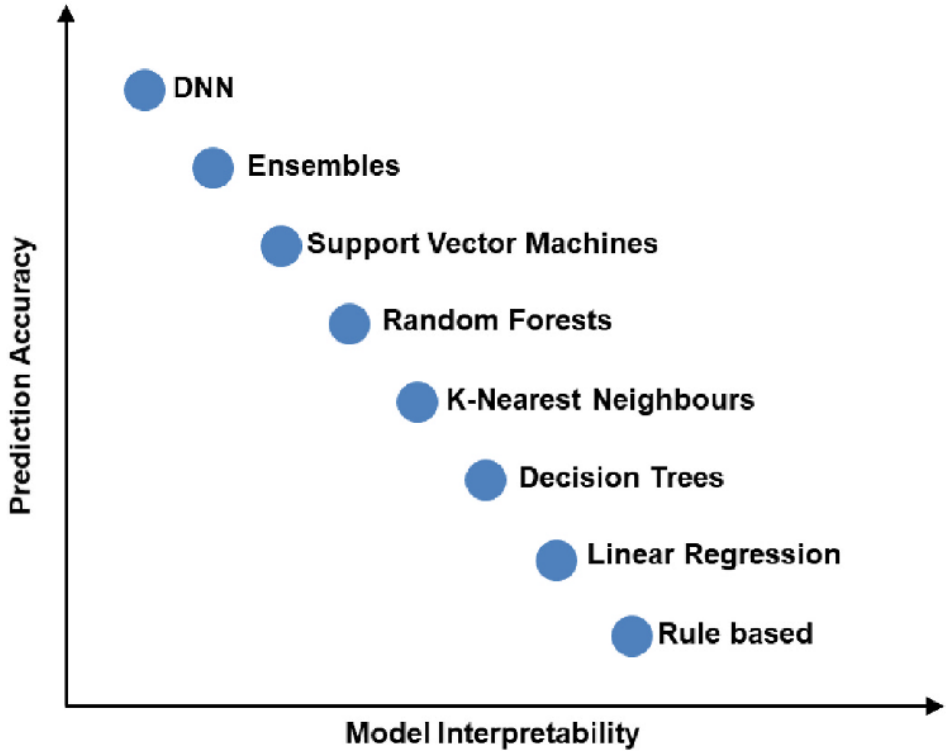

Figure: Explainability of a machine learning model is usually inverse to its prediction accuracy - the higher the prediction accuracy, the lower the model explainability

Two Categories of Explainable AI

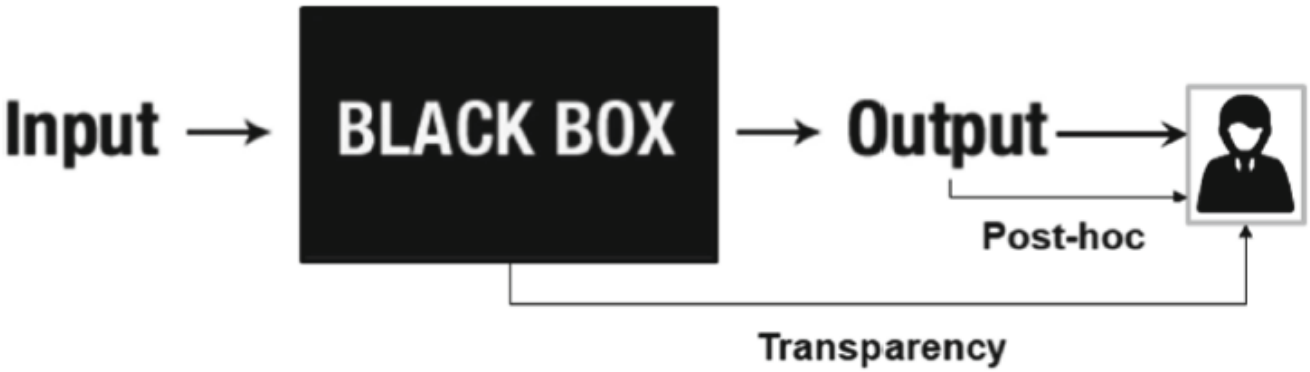

Figure: Two categories of Explainable AI work: transparency design and post-hoc explanation.

Two main strands of explainable AI are:-

- Transparency design
- Post-hoc explanation

Scope of Interpretability

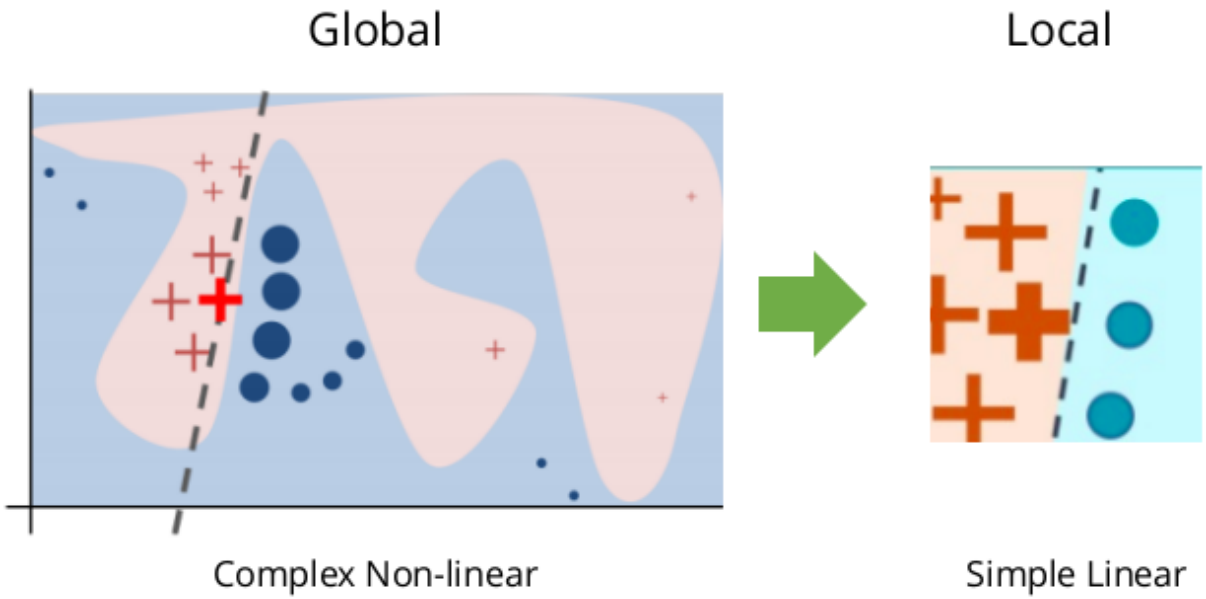

Figure: The Scope of Interpretability

Model Awareness

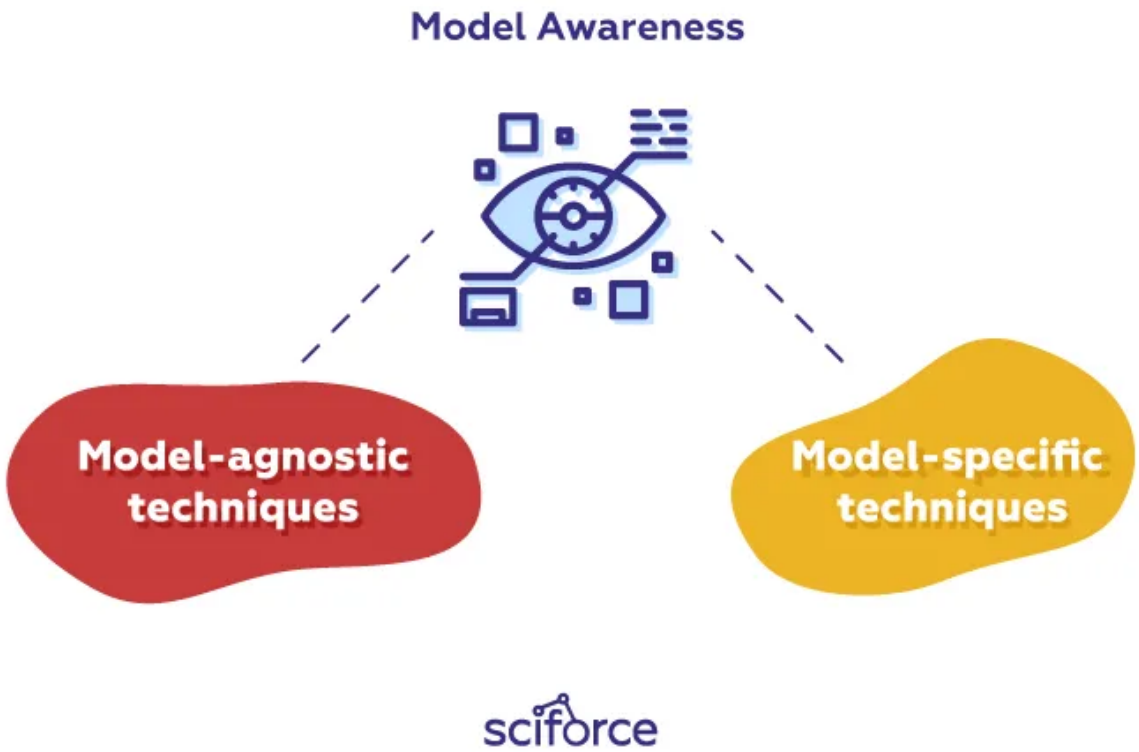

Figure: The applicability of XAI Methods

Importance of XAI

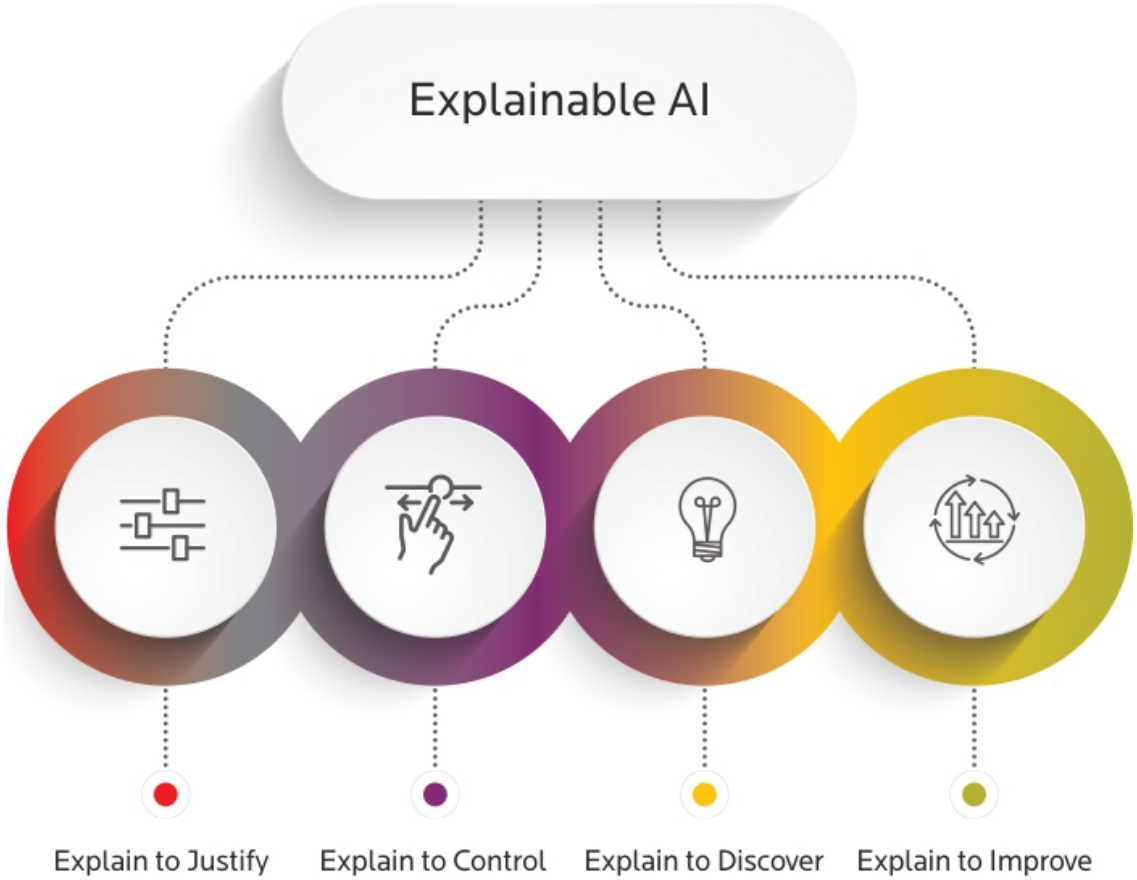

Figure: Why do we need to explain our model

Importance of XAI

- Explainable AI is important to the users who utilize the AI system.

Importance of XAI

- Explainable AI is important to the users who utilize the AI system.
- Explainable AI is important to the people who are affected by AI decision.

Importance of XAI

- Explainable AI is important to the users who utilize the AI system.
- Explainable AI is important to the people who are affected by AI decision.
- Explainable AI could help developers to improve AI algorithm.

Challenges and Future Directions

- Need of a more trustworthy and transparent AI
- Goal is to produce "glass-box" models.

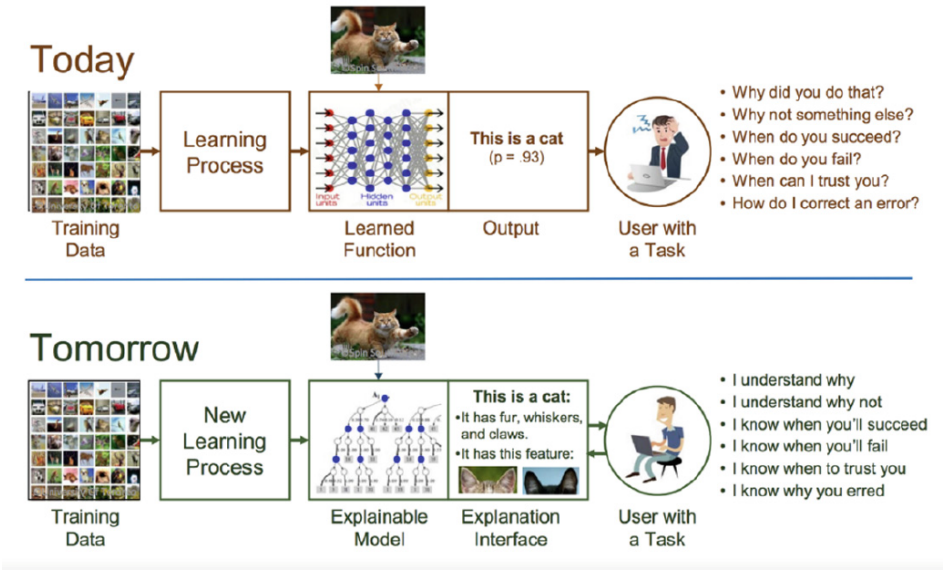

Figure: Explainable AI (XAI) Concept expected in Future.

Challenges and Future Directions

- Humans require explicit knowledge to explain and understand.
- DNN acquire and use implicit knowledge in the form of probabilistic models.
- Other AI methods model explicit knowledge, such as Knowledge Graphs
- Efforts are made to bring these two different worlds together.

References

- https://medium.com
- https://www.ncbi.nlm.nih.gov/pmc/articles
- Explainable AI: A brief survey on history, research areas, approaches and challenges
- Explainable AI - A Brief Overview
- https://ieeexplore.ieee.org/abstract/document/8490530
- https://www.techtarget.com/whatis/definition/explainable-AI-XAI
- https://insights.sei.cmu.edu/blog/what-is-explainable-ai/
- Interpretable Machine Learning by Christpher Molnar