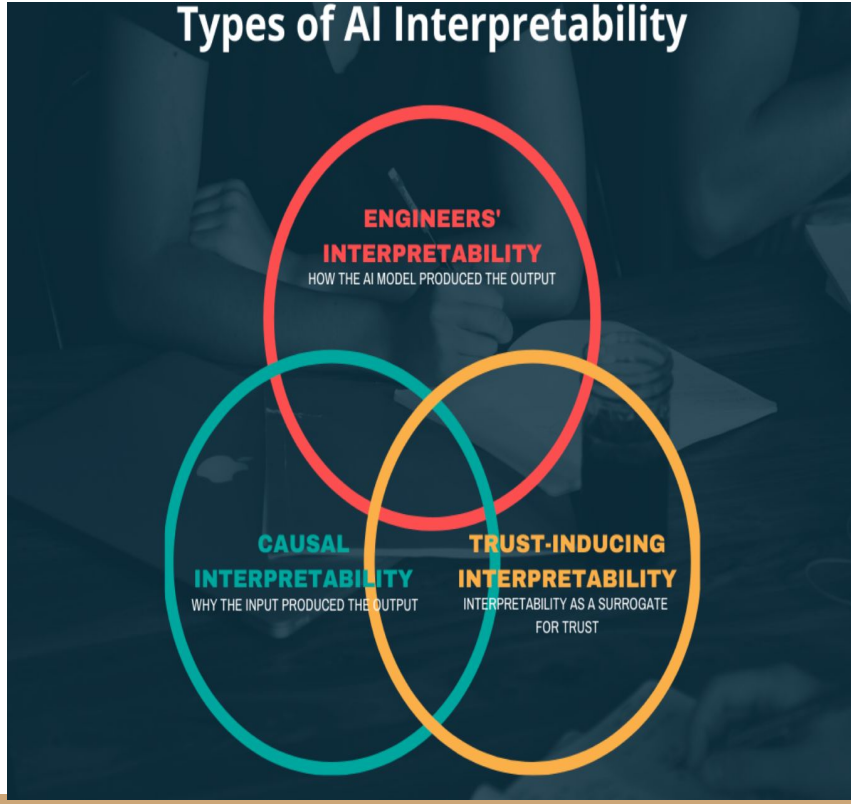
Understanding and Introspecting AI

Contents

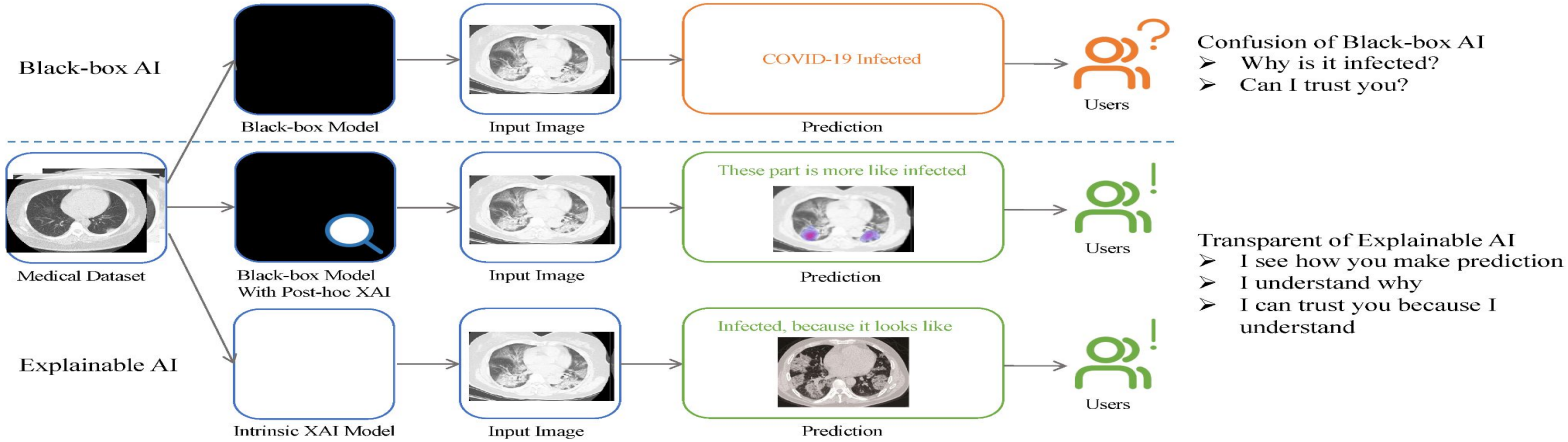
Interpretable AI

- It refers to model transparency and understand exactly why and how the model is generating predictions.
- We need to observe the inner mechanics of the AI/ML method used.
- This leads to interpreting the model's weights and features to determine the given output.



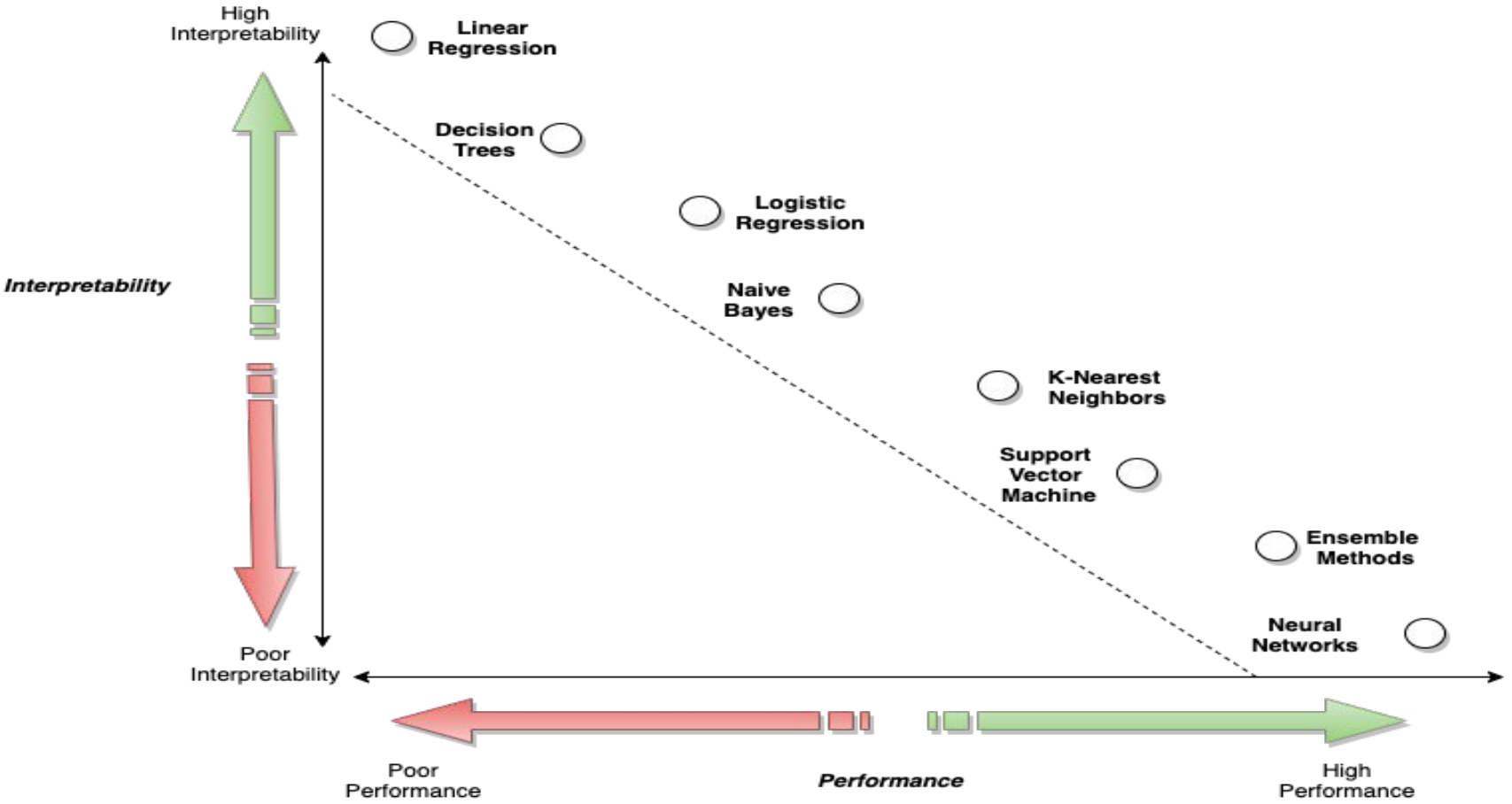
Explainable AI

- Refers to the capability of AI systems to provide understandable explanations for their outputs or recommendations.
- It addresses the inherent opacity of machine learning models, enabling stakeholders to comprehend why a specific decision was made or action was recommended.
- Once a model is Interpretable then it can be Explainable by techniques such as feature importance analysis, decision trees, and model-agnostic approaches like LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive explanations).



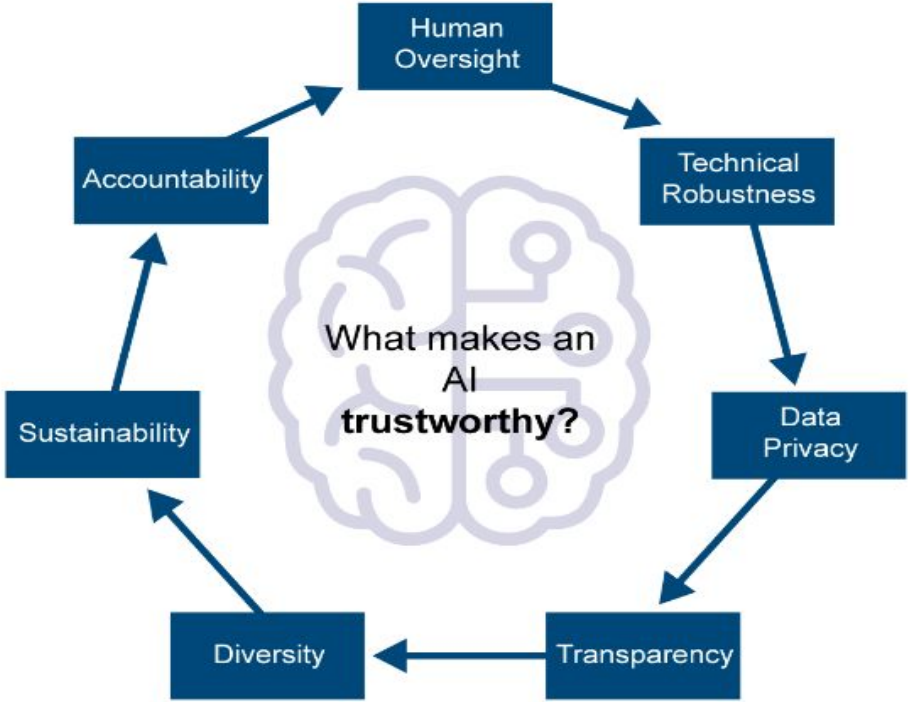
	Interpretable AI	Explainable AI
Focus	Concentrates on making the AI system's internal processes understandable.	Concentrates on providing explicit justifications and reasons for the model’s outputs.
Accessibility	More concerned with the overall understandability of the AI model.	Specifically targets the clarity of explanations provided for individual decisions made by the model.
User Perspective	User-centric, aiming to ensure that users can comprehend the system's functioning.	User-focused, with a primary goal of delivering meaningful explanations to users about the AI's decisions.
Implementation	Involves in simpler model architectures or visualizations that make it easier to follow the decision process.	Often incorporates techniques and methods designed explicitly for generating human-understandable justifications for AI decisions. (LIME)

Interpretability & Performance of ML Algorithms



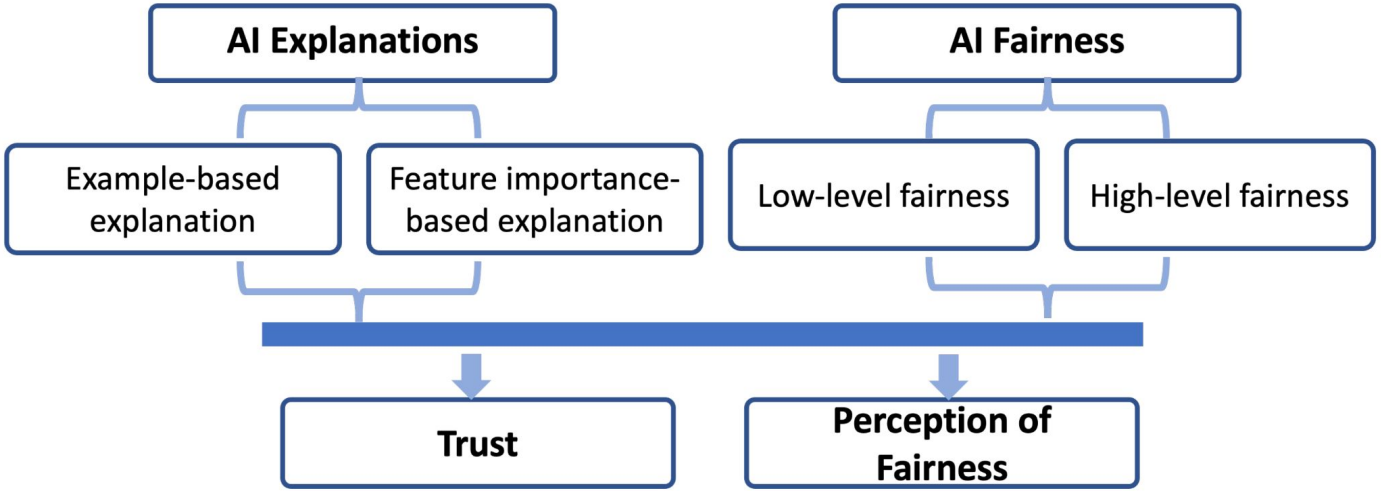
Trustworthy AI

- AI is considered trustworthy when its outputs can be easily interpreted and explained, fostering a clear understanding of the model's decisions.
- Trustworthy AI instills confidence by providing transparent, reliable, and comprehensible results, instilling faith in the system's reliability and accountability.



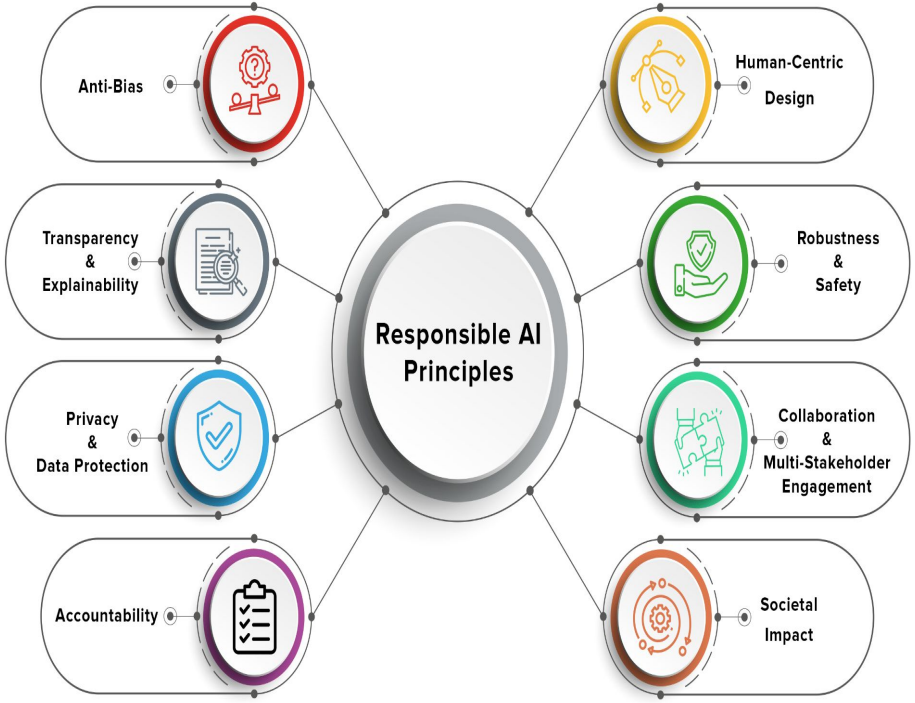
Fairness AI

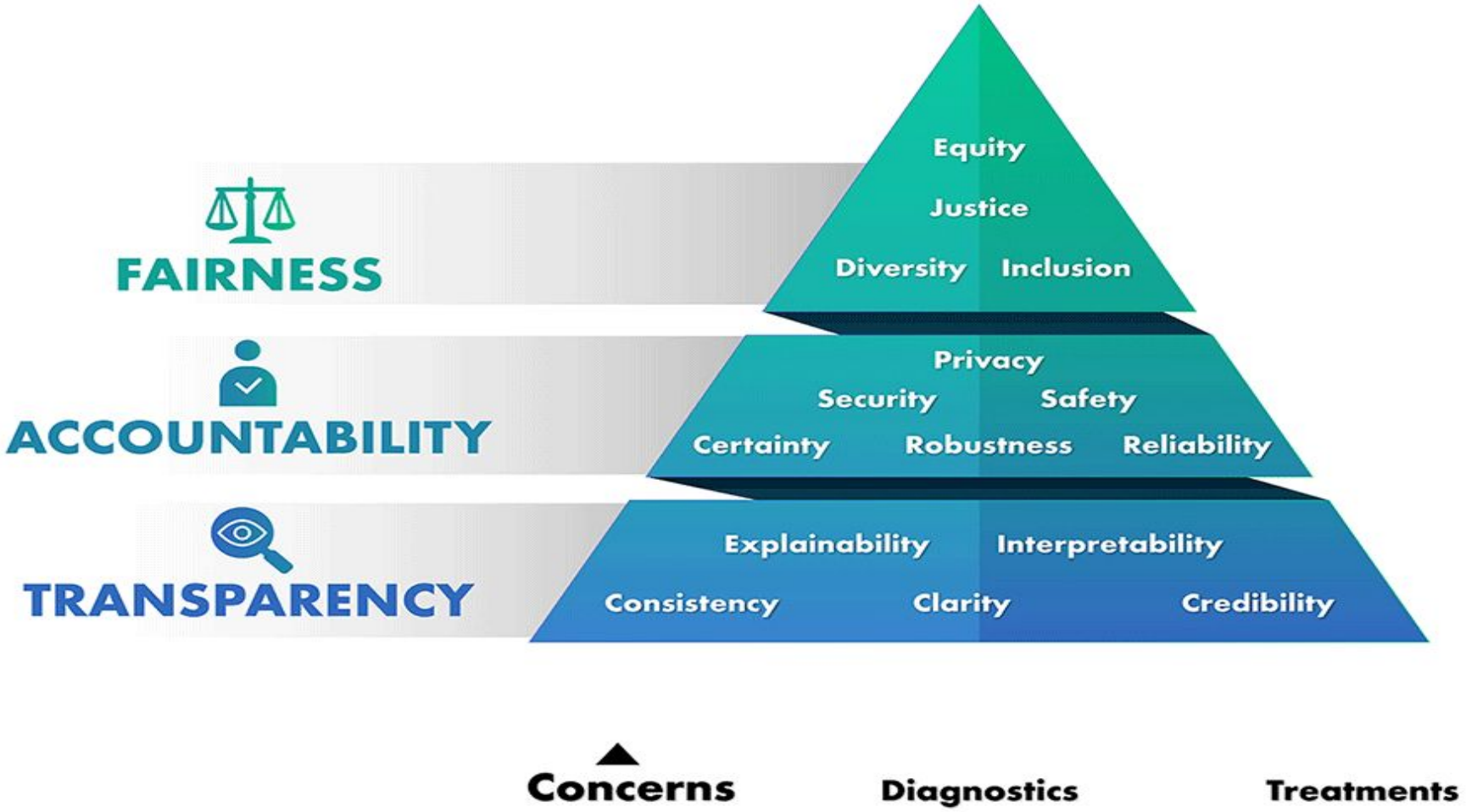
- The goal of fairness in AI is to ensure that AI systems make decisions that are unbiased, equitable, and do not propagate or reinforce societal disparities.
- Fairness metrics are tools used to measure and mitigate bias in AI systems.



Responsible AI

- It is the practice of designing, developing, and deploying AI with good intention to empower employees and businesses, and fairly impact customers and society—allowing companies to trust and scale AI with confidence. Artificial Intelligence.




FAIRNESS
ACCOUNTABILITY
TRANSPARENCY
Equity
Justice
Diversity
Inclusion
Privacy
Security
Safety
Certainty
Robustness
Reliability
Explainability
Interpretability
Consistency
Clarity
Credibility
Concerns
Diagnostics
Treatments