THE PLACE OF USEFUL LEARNING

University of Strathclyde Engineering

From Black Box to Glass Box: A Review into Making Artificial Intelligence Explainable

EE986 - Group 8

D. Das, G. Lundy, S. Nikolaou, N. Patel

MSc. Machine Learning and Deep Learning 2022-23

What is Explainable Artificial Intelligence?

Black Box Algorithm → Transparency / XAI Tools / Traceability → Transparent and Explainable

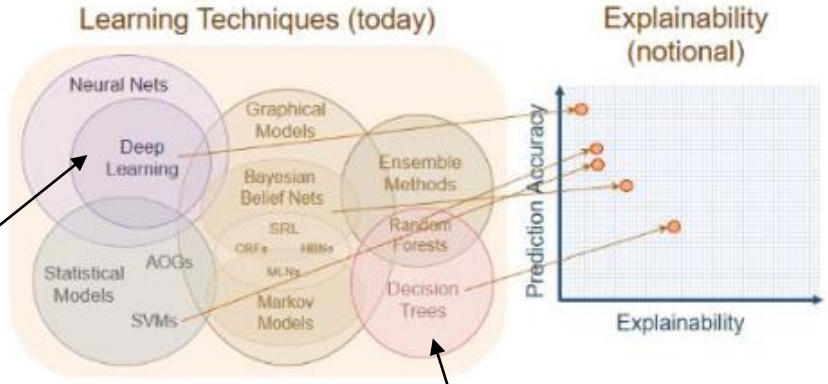
Each individual component of the AI model should be able to provide an explanation, not only the learning algorithm.

Feature Maps	Reoccurring Activation Patterns
- CNN	- RNN

Feature importance is used for:

- Mitigating bias
- Highlighting further Research
- Increasing accuracy
- Reducing Learning Costs

The explainability of a machine is typically the inverse of its prediction accuracy.

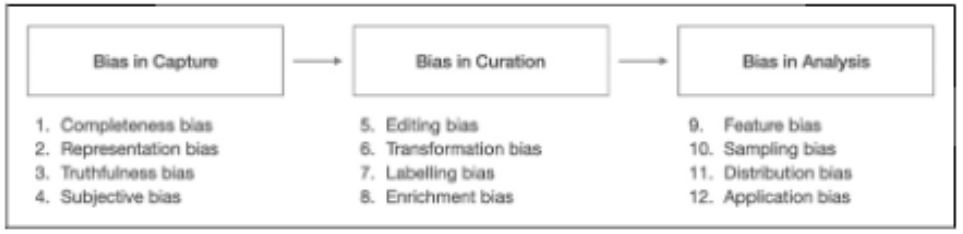


DNNs are extremely precise, but without XAI Tools, they are complete black boxes with no explanation possible.

Decision trees have a very high level of explainability but have the lowest shown prediction accuracy.

Ethical AI

Across all sectors, there is a growing need for ethical, explainable AI.



Types of bias outlined by S. Richardson

Protected characteristics should not be allowed to influence the model unless it is directly involved in the explicit purpose of the model.

Principles established by the Belmont Report for Behavioural Sciences written in 1978 as a starting point for ensuring ethical Human-AI interactions:

- The personal autonomy of a person should not be violated (maintain free-will when interacting with the technology).
- The benefits brought about by the development of technology must outweigh the negatives.
- The benefits and risks must be distributed equally across all peoples using the technology (no person should be discriminated against based on their personal background such as race, gender, and religion).

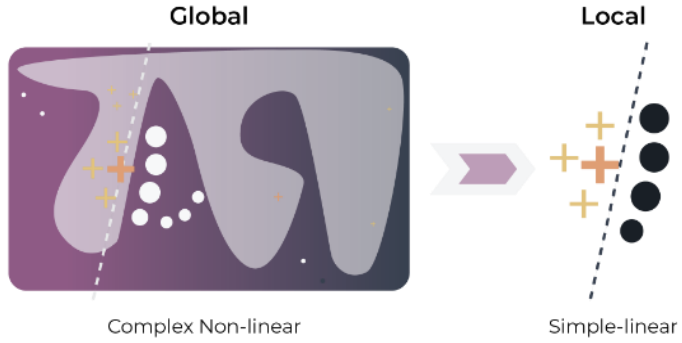
State Of The Art

Goals

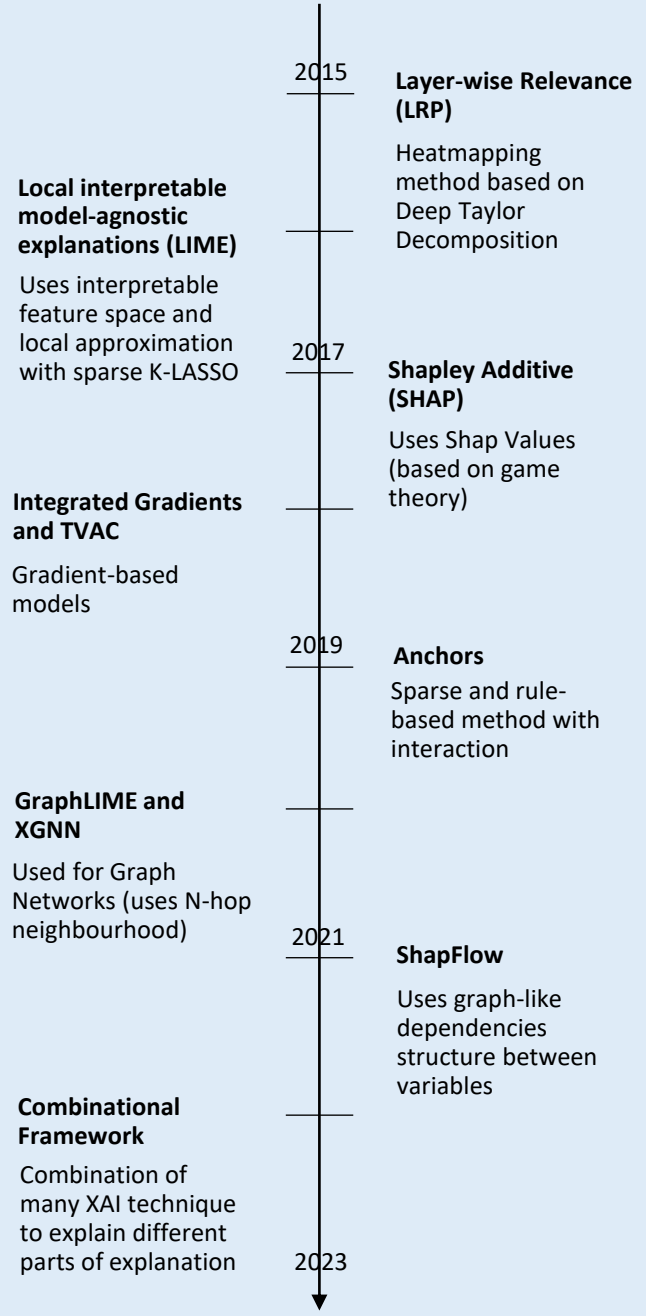
Trust | Understandability | Fairness from Bias | Insightfulness | Causality | Transferability | Data Privacy

Steps for Goals Setting:

1. What is the Problem?
   - Regression, Classification, Data Generation
2. Comprehend the Data and the Data Type
3. Planned Methodology
4. Evaluation and Reliability



Source: https://arize.com/glossary/local-interpretable-model-agnostic-explanations-lime/

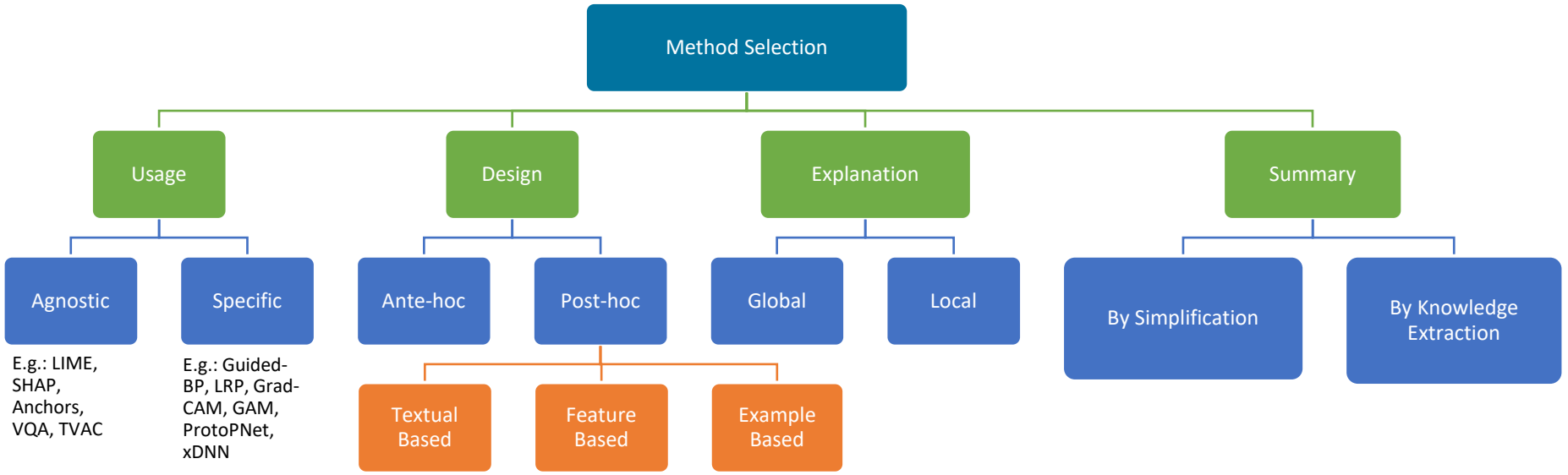


TIMELINE OF PROGRESSION

State Of The Art

Taxonomy (Methodologies and Evaluations)

- 3 Stages of Transparency: **Simulatability** | **Intelligibility** | **Algorithmic Transparency**
- For complex problems, Post-Hoc is preferable to Ante-Hoc (explanation for the black box)
- Complexity depends on Target Users: **AI Experts** | **Domain Experts** | **AI Laymen**
- Evaluation Metric: **Human-Centric** | **Computer Centred**



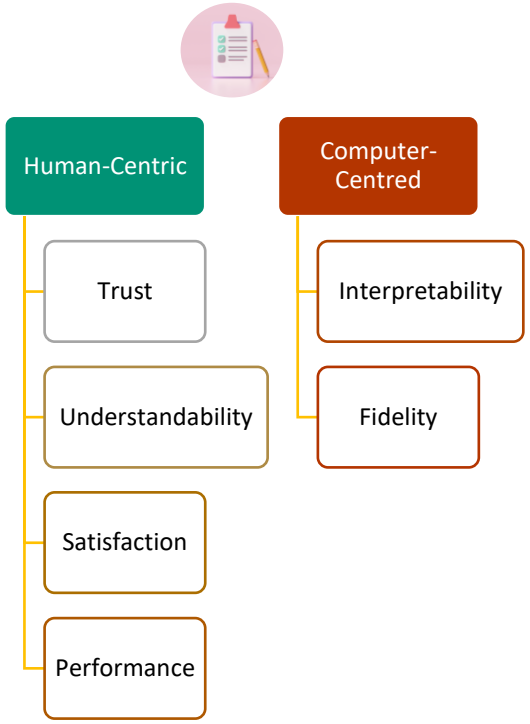


Diagram Reference: I. Ahmed, G. Jeon and F. Piccialli, "From Artificial Intelligence to Explainable Artificial Intelligence in Industry 4.0: A Survey on What, How, and Where," in IEEE Transactions on Industrial Informatics, vol. 18, no. 8, pp. 5031-5042, Aug. 2022, doi: 10.1109/TII.2022.3146552.

XAI Applications

- Transparency and trust issues lead to reluctance in incorporating AI in core functions – loss of human lives
- XAI can build up trust

XAI in Manufacturing

- Can help improve quality control, optimise production processes, and reduce costs
- Predictive maintenance – predict when machines are likely to fail and recommend maintenance
- Can help improve product quality by detecting defects – very important as significant amounts of time are spent for manual inspection

XAI Applications

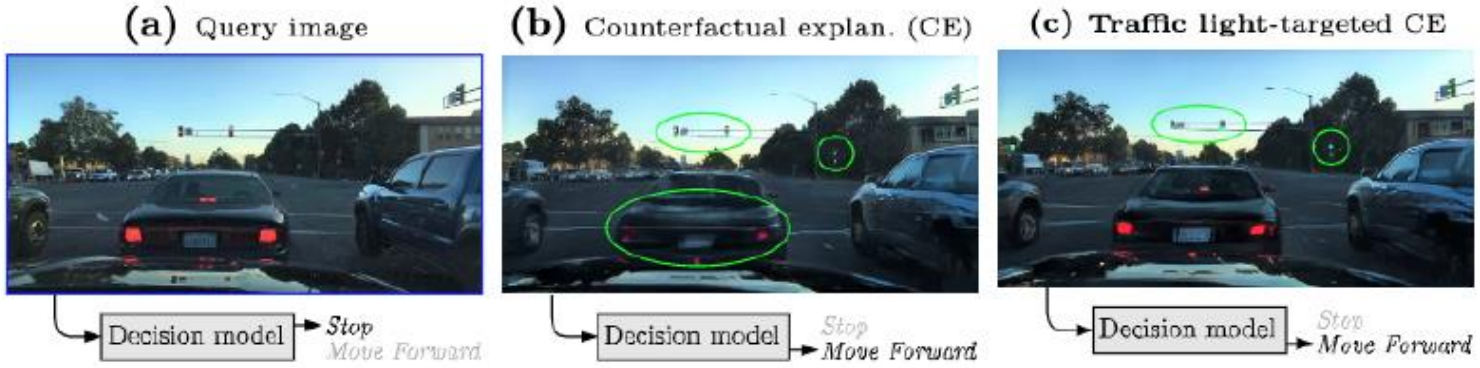
XAI in Healthcare

- AI has expanded in Clinical Decision Support, Drug Discovery, Hospital Management, Predictive Medicine and Patient Data
- Explainable models can help doctors:
  - Understand how AI models are making decisions
  - Identify personalised treatments based on relevant factors
- Medical image classification – highlight most influential parts
- X-rays to distinguish COVID-19 pneumonia, and images to detect different types of cancer
- Hospital Management – identify patients at risk of readmission

XAI Applications

XAI in Autonomous Driving

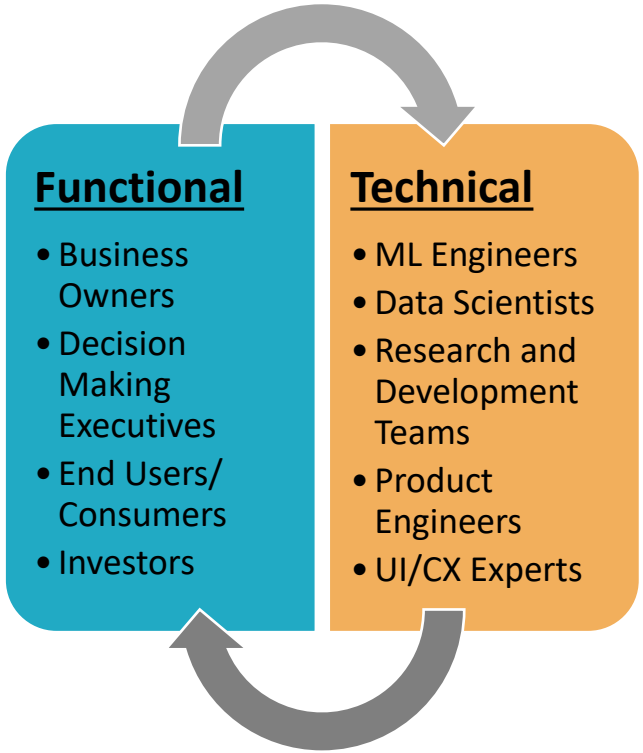
- Crucial for ensuring safety and building trust between people and cars
- Can help explain why an AV made a decision/took action
- It is of supreme importance to know what caused the above
- Counterfactual explanations for content-based insights
- XAI can also help identify potential risks and prevent accidents



Source: P. Jacob, É. Zablocki, H. Ben-Younes, M. Chen, P. Pérez, and M. Cord, "STEEX: Steering Counterfactual Explanations with Semantics", arXiv:2111.09094 [cs], Jul. 2022, Available: https://arxiv.org/abs/2111.09094

Business Implementation

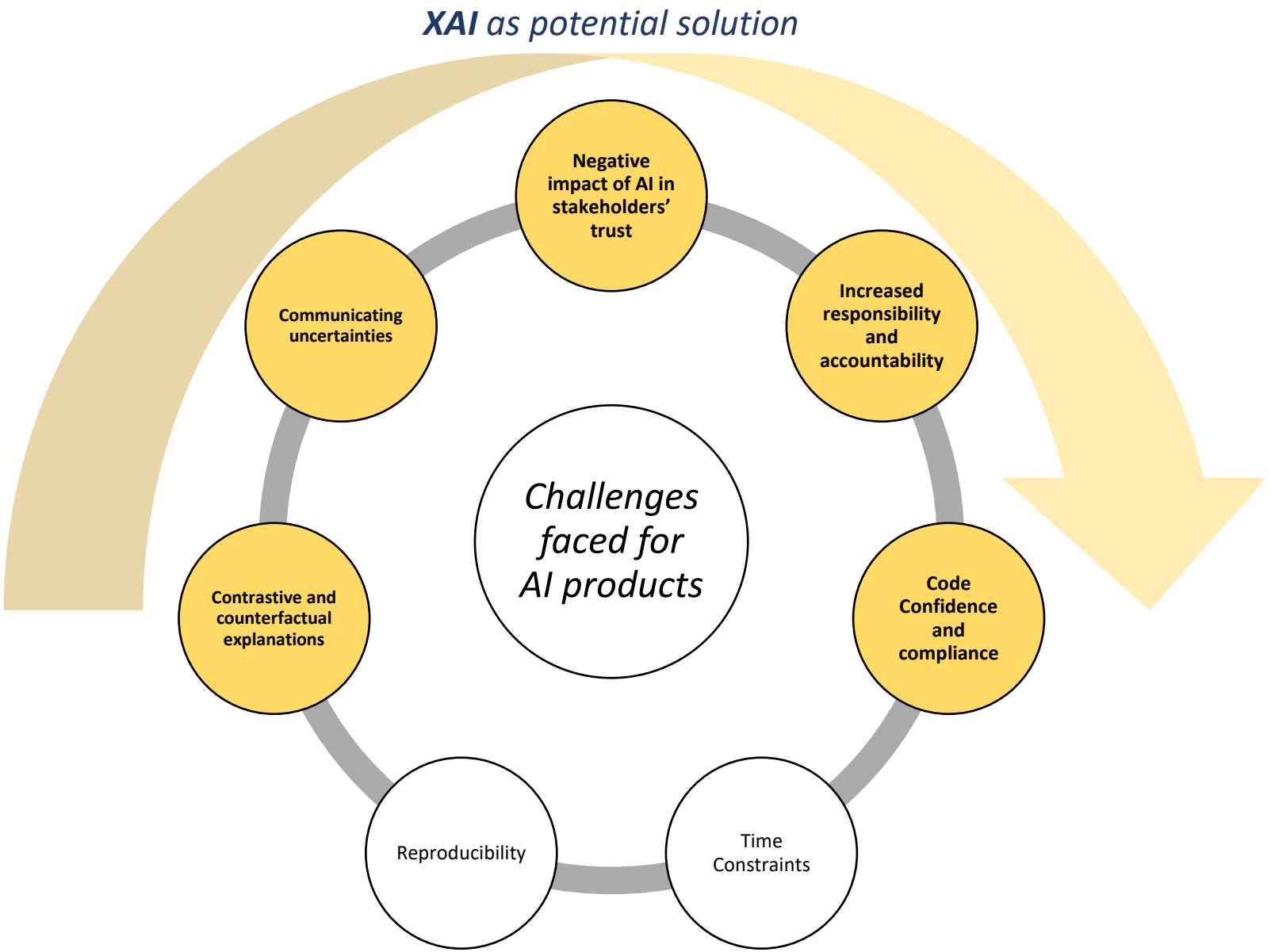
Bridging the gap between Technical and Functional Stakeholders



- Conventional ways such as ‘open sourcing’ products and their features may not work for black box algorithms
- XAI to help to both groups in business implementation by
  - **Adding Explainability and Trustworthiness**
  - **Adding Debugging, Monitoring and Regulation capabilities**
- Two fold AI implementations:
  AI in Core Products + Explainability solutions
- Proposed method: Adopting more connected desiderata definitions and advanced supporting documentations i.e. increased efforts overall

Image Source: DALL·E 2 by OpenAI

The Future of XAI



Total estimated AI opportunity of $15 Trillion* *due to AI driven transformations and universal embedment of AI within product components*

**According to pWc XAI report: https://www.pwc.co.uk/xai*