

Introduction to Computational Intelligence



Explainable Artificial Intelligence

Igor Farkaš
 Centre for Cognitive Science
 DAI FMFI Comenius University in Bratislava

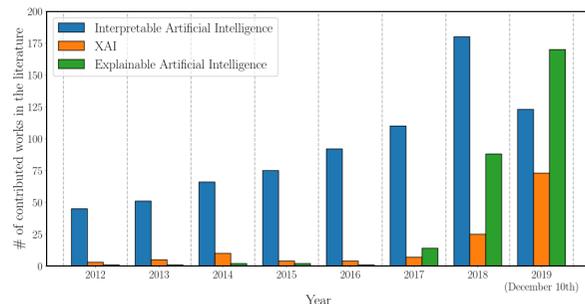
Based on Barredo Arrieta et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115

Shapshots of AI history

- 1950s – Alan Turing, Norbert Wiener
 - small cybernetic robots (animats) ← behaviorism
 - e.g. turtles (Walter), mouse (Shannon) ← NNs, RL involved
- 1955 – Simon, Newell, Minsky, McCarthy – birth of AI (cognitivism)
 - focus on human mental tasks: math. theorem proving, problem solving, board games, later extended: e.g. medical diagnosis, legal argumentation, NLP, common sense reasoning
 - based on symbolic AI (as opposed to numeric AI = NNs), mostly transparent
- 1986 – revival of NNs (thank to backpropagation)
- 2010 – outburst of deep NN (thank to fast computers and big data), seen as black-boxes (with tons of parameters)

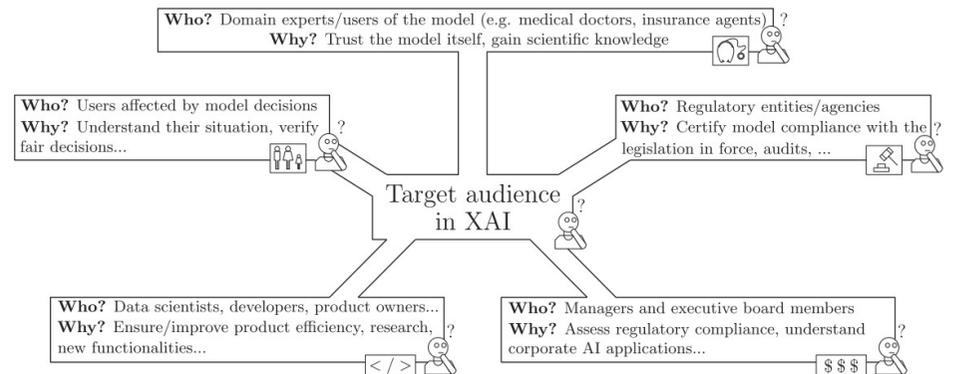
Introduction

- Current hype in narrow AI thank to flourishing deep learning
 - DL started after 2006
- growing applications of DL in various practical domains
- But, neural networks as black-box models
- => need for interpretable, transparent, explainable AI (XAI)



Different purposes of explainability in ML

Two goals dominant: need for model understanding, and regulatory compliance



(Rossi, 2019)

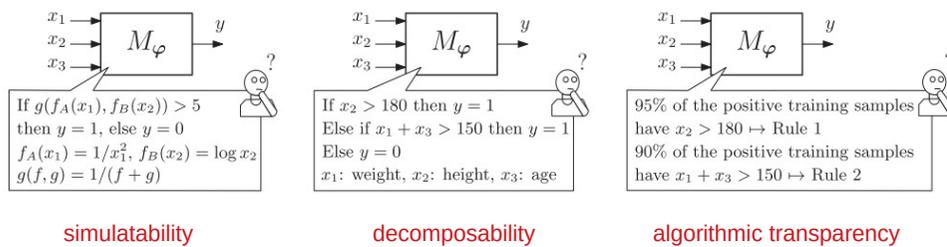
Terminology

- **Understandability (intelligibility)** – of a model function by a human, without any need for explaining its internal structure or the algorithmic means by which the model processes data internally.
- **Comprehensibility** – refers to the ability of a learning algorithm to represent its learned knowledge in a human understandable fashion.
- **Interpretability** – the ability to provide the meaning in understandable terms to a human.
- **Explainability** – an interface between humans and a decision maker, i.e., at the same time, both an accurate proxy of the decision maker and comprehensible to humans.
- **Transparency** – Since a model can feature different degrees of understandability, transparent models fall into three categories: simulatable models, decomposable models and algorithmically transparent models.

Goals towards explainability

Goal	Target audience
Trustworthiness	Domain experts, users of the model affected by decisions
Causality	Domain experts, managers and executive board members, regulatory entities/agencies
Transferability	Domain experts, data scientists
Informativeness	All
Confidence	Domain experts, developers, managers, regulatory entities/agencies
Fairness	Users affected by model decisions, regulatory entities/agencies
Accessibility	Product owners, managers, users affected by model decisions
Interactivity	Domain experts, users affected by model decisions
Privacy awareness	Users affected by model decisions, regulatory entities/agencies

Different levels of transparency

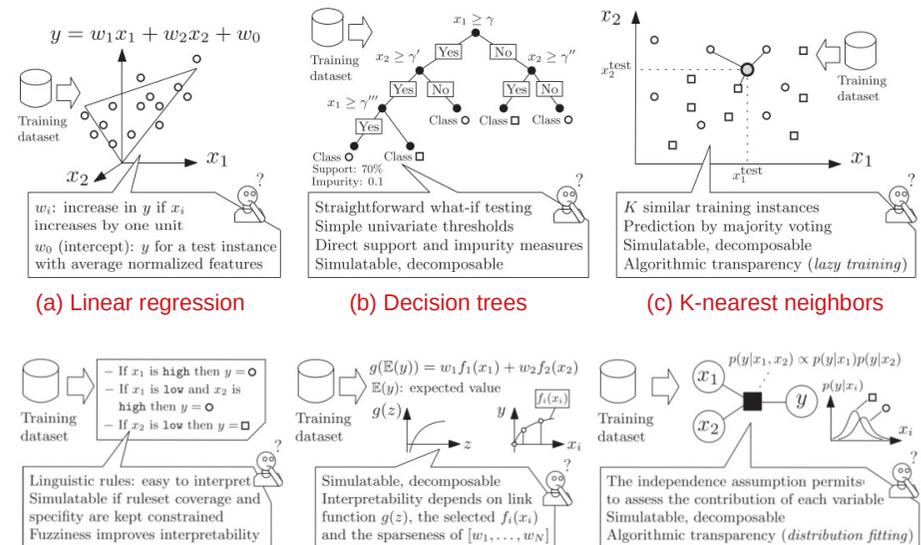


Enable mental simulation, e.g. a simple linear model (a simple perceptron), but not a symbolic system with too many rules.

Every input must be readily interpretable. Every part of the model must be understandable by a human without the need for additional tools.

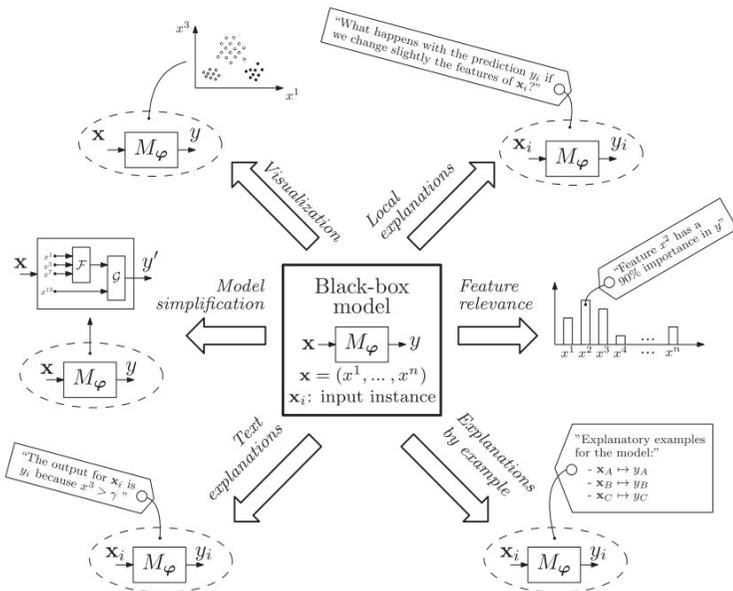
Model has to be fully explorable by means of mathematical analysis and methods (e.g. linear vs deep NN).

Levels of transparency of different ML models



(a) Linear regression (b) Decision trees (c) K-nearest neighbors (d) Rule-based learners (e) Generalized additive models (f) Bayesian models

Post-hoc explainability approaches

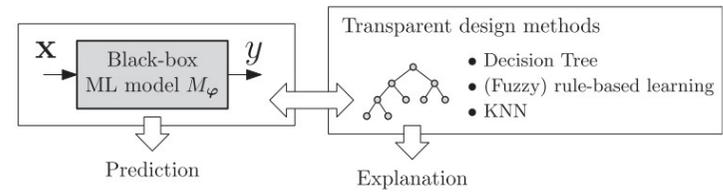


9

Non-transparent approaches

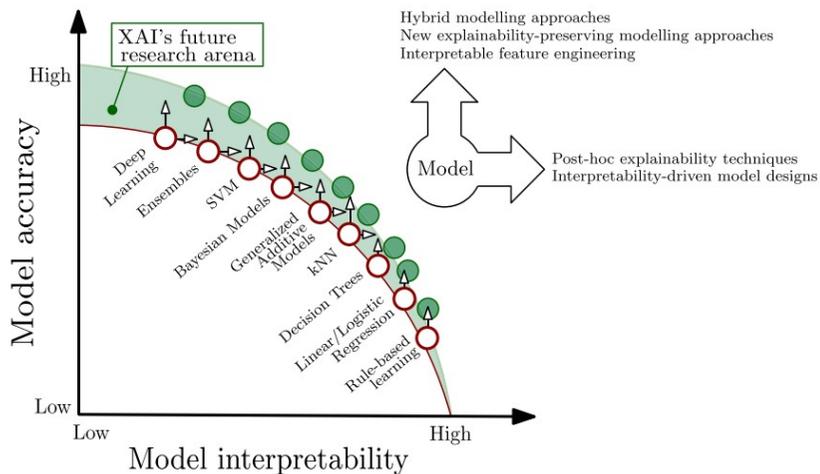
- **model-agnostic** host-hoc explainability, by
 - simplification, feature relevance, local, visualisation
- **model-specific** host-hoc explainability in
 - support vector machines, multilayer, convolutional, recurrent NN, utilize the methods above

Possible approach: Hybrid models



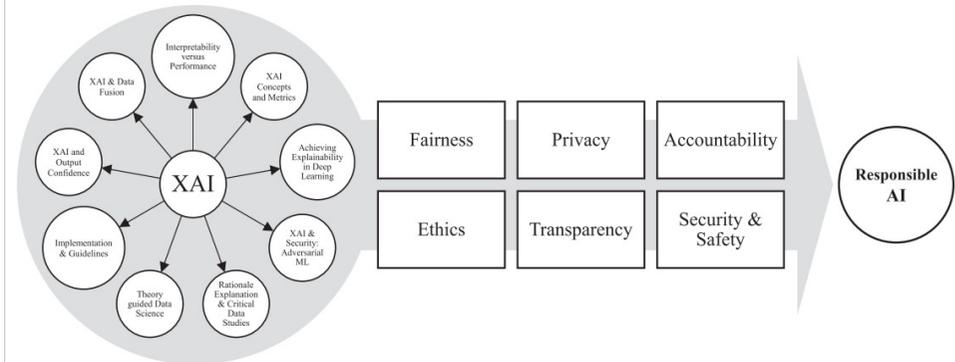
10

Trade-off in AI methods



11

Authors' vision



12