

# Introduction to Explainable AI (XAI)

---



Alexander Gegov

Associate Professor in Computational Intelligence

University of Portsmouth

[alexander.gegov@port.ac.uk](mailto:alexander.gegov@port.ac.uk)

# Presentation Outline

---

1. AI Characteristics
2. AI Bias
3. XAI Taxonomy
4. XAI Types
5. XAI Formats
6. XAI Challenges
7. XAI Solutions
8. XAI Concept
9. XAI Books

# 1. AI Characteristics

---

- Responsibility
- Trustworthiness
- Safety
- Reliability
  
- Fairness
- Transparency
- Interpretability
- Explainability

## 2. AI Bias

---

- Real world bias – reflected in data bias
- Data bias – encoded in algorithmic bias
- Algorithmic bias – uncovered by XAI
- XAI – used to mitigate real world bias

### 3. XAI Taxonomy

---

- Dataset application scope  
(global, local)
- Machine learning models  
(interpretable, black box)
- Model explanation methods  
(specific, agnostic)

## 4. XAI Types

---

- Pre-model

(data – XAI – black box model – user)

- In-model

(data – interpretable XAI model – user)

- Post-model

- (data – black box model – XAI – user)

## 5. XAI Formats

---

- What-if
- Counterfactual
  
- Example based
- Constructive

## 6. XAI Challenges

---

- Evaluation
- Formalisation
  
- Adoption
- Acceptance
  
- Causality
- Reasoning



## 7. XAI Solutions

---

- Explainable models

  - Global analysis (all inputs and outputs)

  - Local analysis (individual inputs and outputs)

- Meaningful explanations

  - Simple structural models (inputs, outputs)

  - Complex structural models (sub-models, connections)

## 7. XAI Solutions

---

- Structural model presentation

  - Directed graph (1-to-1 mapping of structural model)

  - Graph edges (external inputs and outputs)

  - Graph nodes (sub-models)

  - Graph edges (internal connections)

- Structural model evaluation

  - Grid (horizontal levels and vertical layers)

  - Number (external inputs and outputs)

  - Number (sub-models and internal connections)

## 7. XAI Solutions

- Macro models (flat, black-box, single node)
  - One 4-input-1-output node (level 1, layer 1)
  - Shallow/concise explanations (for expert users)
  - $y = f(x_1, x_2, x_3, x_4)$
- Micro models (hierarchical, white-box, multiple nodes)
  - Two 2-input-1-connection nodes (levels 1-2, layer 1)
  - One 2-connection-1-output node (level 1/2, layer 2)
  - Deep/detailed explanations (for non-expert users)
  - $y = f[f_1(x_1, x_2), f_2(x_3, x_4)]$

## 7. XAI Solutions

---

- Model is less complex than reality  
Flat model for a hierarchical process  
Rough/superficial explanations
- Model is more complex than reality  
Hierarchical model for a flat process  
Detailed/abstract explanations
- Model is as complex as reality  
Flat/hierarchical model for a flat/hierarchical process  
Precise/adequate explanations

## 7. XAI Solutions

---

- Quantitative approach  
Data (objective/observation context)
- Qualitative approach  
Knowledge (subjective/consultation context)
- Hybrid approach  
Data (objective/observation context)  
Knowledge (subjective/consultation context)

## 7. XAI Solutions

---

- Model efficiency

Decreases when MS/ME gets worse (FM)

Increases when MS/ME gets better (HM)

- Model accuracy

Increases when MS/ME gets worse (FM)

Decreases when MS/ME gets better (HM)

(MS – model simplicity, ME – model explainability)

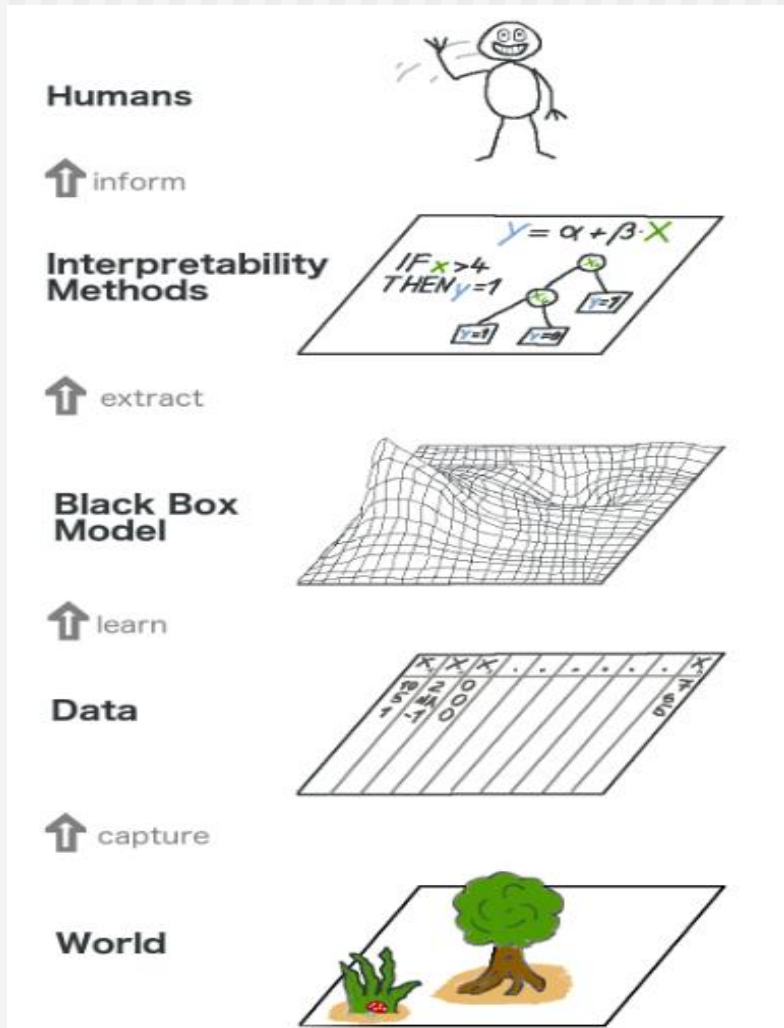
(FM – flat model, HM – hierarchical model)

## 7. XAI Solutions

---

- Mortgage application (flat and hierarchical models)  
Outcome = FM (income, assets)  
Outcome = HM [repayments (income),  
deposit (assets)]
- Job application (flat and hierarchical models)  
Outcome = FM (qualifications, experience)  
Outcome = HM [effectiveness (qualifications),  
efficiency (experience)]

# 8. XAI Context





# 9. XAI Books

