# 📘 Explainable AI – Full Lecture Notes

**Lecturer: Natnael Argaw Wondimu**

**Institution**: School of Information Technology and Engineering, AAiT
**Based on**: Selected Topics in AI

---

## 📍 Chapter 1: Topics in Machine Learning – Interpretability and Explainability

### Motivation for Model Understanding:

- ML is everywhere; decisions affect real lives.
- Understanding helps with:
    - **Debugging**
    - **Bias detection**
    - **Recourse for affected individuals**
    - **Building trust in decisions** (e.g., healthcare)

### Achieving Model Understanding:

1. **Inherently Interpretable Models**
    - Built to be transparent and explainable by design.
    - E.g., decision trees, linear models.
2. **Post-hoc Explanations**
    - Explaining black-box models after training.

- E.g., LIME, SHAP, Visualizations.

### Interpretability vs. Accuracy:

- Trade-off often exists.
- Choose interpretable models if they are sufficiently accurate.
- Else, use post-hoc explanations for black-box models.

---

## 📍 Chapter 2: Defining and Evaluating Interpretability

### Definition:

- **Interpretability**: The ability to present model behavior in understandable human terms.

### When is Interpretability Needed?

- Not all ML systems require it (e.g., ad servers).
- Needed when:
  - Problem formulation is incomplete
  - Systems are deployed in high-stakes domains (ethics, safety)
  - Regulatory or fairness concerns exist

### Motivations (Desiderata):

1. **Trust**: Understanding promotes confidence in systems.
2. **Causality**: Moving beyond correlation to explanation.
3. **Informativeness**: Aids in decision-making beyond raw output.
4. **Fair & Ethical Decisions**: Prevent discrimination (e.g., racial bias).

### Taxonomies of Interpretability:

- Based on **evaluation**, **application**, and **methods** (Doshi-Velez & Kim, 2017).

---

# 📍 Chapter 3: Caveats and Challenges

### Challenges:

1. **Interpretation ≠ Explanation**: Human explanations may not match model logic.
2. **Transparency Issues**:
    - Complex models = hard to interpret (deep learning, ensembles)
    - Trade-off between performance & interpretability
    - Lack of standardization

### Types of Transparency:

- **Simulatability**: Entire model is mentally executable by humans.
- **Decomposability**: Each part (input, parameter, calculation) is understandable.

### Post-hoc Explanation Types:

1. **Text explanations**: Natural language justifications.
2. **Visualization**: E.g., t-SNE, saliency maps.
3. **Example-based**: k-nearest neighbors reasoning.
4. **Local approximations**: LIME creates interpretable models around specific predictions.

### Warnings:

- Post-hoc explanations can be **plausible but misleading**.

- **Selective transparency** can be harmful or discriminatory.
- More information isn't always better (Braess' paradox).

---

# 📍 Chapter 4: Human Factors in Explainability

## Interpretability Stakeholders:

- **Model Developers**: Debugging, understanding.
- **Domain Experts**: Trust, knowledge incorporation.
- **Auditors/Regulators**: Compliance, fairness.
- **End Users**: Control and personalization.
- **Society**: Ethics and public understanding.

## Practitioner Needs:

- Interpretability tools must be:
    - **Cooperative** (support domain knowledge)
    - **Process-based** (support iterative refinement)
    - **Context-dependent** (tailored to user/domain)
    - **Aligned with mental models** (human-machine understanding bridge)

## Methodology:

- Need for accessible, repeatable methods tailored to real needs.
- Current tools often **overtrusted** or **misused**.

---

# 📍 Chapter 5: Evaluating Interpretability (Human Evaluation)

### Goal:

- Evaluate interpretability using **human judgment**.

### Method:

- Controlled experiments using:
    - **Feature Importance**
    - **Rule Extraction**
    - **Partial Dependence Plots**

### Findings:

- Human understanding improves with interpretability tools.
- **Feature importance** is especially effective.
- Visual tools help relate inputs to outputs.

### Implications:

- Positive trust effects.
- Highlights need for **human-centered evaluation** in tool development.

---

# 📍 Chapter 6: Rule-Based Interpretability Approaches

### 1. Interpretable Rule Lists (IRL)

- Developed by **Letham et al.**

- Uses a **greedy algorithm** to build a rule list balancing accuracy & interpretability.
- Each rule: IF [condition] → THEN [prediction]
- Simplicity enforced (e.g., limit rule length).

**Pros:**

- Concise and easy to understand.
- Balances performance and clarity.

**Cons:**

- Struggles with complex decision boundaries.
- Limited feature interaction.

---

## 2. Interpretable Rule Sets (IRS)

- Developed by **Lakkaraju et al.**
- Rule induction + pruning:
    - Extracts rules using thresholds & logic.
    - Removes redundant/irrelevant rules.

**Pros:**

- Richer representation of decision space.
- Customizable pruning for better interpretability.

**Cons:**

- More complex than rule lists.

- Risk of pruning essential rules.

---

## 📍 Seminar Topics Overview

(Assigned separately — for team presentations)

- Rule-Based Approaches
- Counterfactual Explanations
- Attention & Concept-based Explanations
- Interactive and Data Attribution Explanations
- Interpreting Generative Models
- Explainability for Fairness
- Mechanistic Interpretability
- Adversarial Attacks vs. XAI
- Systematic Review of Trustworthy XAI

---

## ✅ Final Remarks

- Choose interpretable models when possible.
- Use post-hoc tools wisely; they don’t open the black box, they approximate.
- Interpretability must be **practical**, **human-centered**, and **ethical**.
- Future work needed in **standardization**, **evaluation frameworks**, and **cooperative tools**.

# 📚 Explainable AI (XAI) — Textual Summary

## 🎯 1. Key Definitions

- **Interpretability**: Explaining model behavior in human-understandable terms.
- **Explainability**: Broader strategies to communicate model decisions (includes post-hoc methods).
- **Transparency**: Clear understanding of how a model works (e.g., parameters, rules).

---

## 🧰 2. Approaches to Explainability

### A. Inherently Interpretable Models

- Built to be understandable from the start.
- Examples: Linear models, decision trees, rule lists/sets.

### B. Post-hoc Explanations

- Applied after training a black-box model.
- Techniques: LIME, SHAP, visualizations, example-based, local approximations.

---

## 🎓 3. Motivations for XAI

- **Trust**: Build user confidence.
- **Causality**: Go beyond correlation.

- **Fairness**: Detect and avoid discrimination.
- **Safety**: Prevent harmful errors in critical domains (e.g., healthcare).

---

## ⚖️ 4. Trade-offs & Challenges

- **Accuracy vs Interpretability**: Simple models may lose performance.
- **Transparency vs Complexity**: Deep models are harder to interpret.
- **Standardization Gap**: Few benchmarks or guidelines for evaluating interpretability.
- **Misleading Post-hoc**: Explanations can sound plausible but be false.

---

## 🔧 5. XAI Techniques and Tools

| Tool/ Method | Type | Description |
|---|---|---|
| **LIME** | Post-hoc | Explains predictions locally |
| **SHAP** | Post-hoc | Feature contribution scores |
| **Rule Lists** | Interpretable | Simple IF-THEN rules (Letham et al.) |

| | | |
|---|---|---|
| **Rule Sets** | Interpretable | Broader, pruned rule groups |
| **t-SNE** | Visualization | 2D plots of high-dimensional data |

---

## 👥 6. Human Factors in XAI

- **Roles**: Developers (debug), Experts (validate), Auditors (assess fairness), Users (understand), Society (ethics).
- **Needs**:
    - Tools that support **cooperation** and **domain knowledge**.
    - Interpretability as a **process**, not a one-time output.
    - Align with **mental models** and **context-specific needs**.

---

## 🧪 7. Evaluation of Interpretability

- Conducted with human subjects (Lage et al.).
- Techniques like **feature importance** and **rule extraction** helped users understand models.
- Emphasis on **human-centered evaluation** and trust-building.

---

## 📋 8. Summary of Rule-Based Methods

### A. Interpretable Rule Lists (IRL)

- Built greedily; concise, human-readable.
- Pros: Simplicity.
- Cons: Limited feature interaction.

### B. Interpretable Rule Sets (IRS)

- Extract → prune rules.
- Pros: Richer decision boundary representation.
- Cons: More complex, pruning can lose info.