Module 21: Transparency and Interpretability

Learning outcomes:

1. Evaluate situations for potential bias in data sets and models.
2. Develop datasheets to document data sets for transparency.
3. Analyse model cards to evaluate transparency and potential bias in models.
4. Analyse the trade-offs of explainability and interpretability in relation to fairness, accuracy and complexity when applying ML models in practice.

Why transparency and interpretability matter

- ML/AI models often function as 'black boxes'.
- Transparency and interpretability make these systems trustworthy, accountable and fair.
- Transparency and interpretability reduce unintended harm and ensure decisions align with human values.

Bias in data

- Bias arises when data reflects historical imbalances, stereotypes or narrow sources.
- Consequences are flawed predictions, discrimination and unfair outcomes.
- Datasheet for data sets:
    - Captures motivation, composition, collection process, preprocessing, uses, distribution and maintenance
    - Promotes transparency, reproducibility and ethical use

Bias in models

- Even with fair data, design and training choices introduce bias.
- Opaque models may disadvantage groups unintentionally.
- Model card:
    - Documents purpose, intended use, limitations, training data, evaluation metrics and ethical considerations
    - Helps stakeholders decide when models can be trusted

Explainability vs interpretability

- **Interpretability:** a human can directly understand how a model works (e.g. linear regression or decision trees).
- **Explainability:** tools (e.g. SHAP, LIME and attention visualisations) generate explanations for complex, black-box models.

- **Regulation:** GDPR grants a ‘right to explanation’ for significant automated decisions.
- **Implications:**
  - Trust: clear reasoning fosters user trust.
  - Accountability: enables auditing and challenging unfair outcomes.
  - Compliance: satisfies legal and ethical obligations.

Trade-offs to consider

- **Interpretability vs accuracy:** simpler models may be less accurate, and complex models may be opaque.
- **Fairness vs complexity:** sometimes reducing complexity increases transparency but decreases fairness.
- **Performance vs accountability:** balance technical optimisation with ethical responsibility.