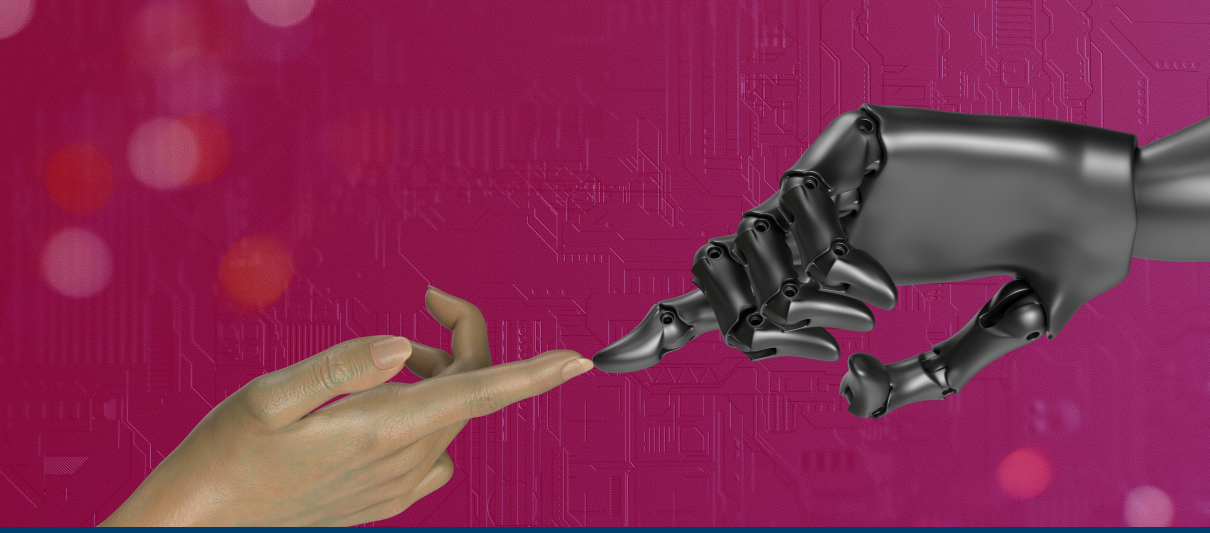


Hands-On Machine Learning

Dr. Derek Bridge | 2026/2027

UCC Coláiste na hOllscoile Corcaigh
University College Cork, Ireland

School of Computer Science and Information Technology
Scoil na Ríomheolaíochta agus na Teicneolaíochta Faisnéise

20. Explainable Artificial Intelligence (XAI)

Explainable Artificial Intelligence (XAI)

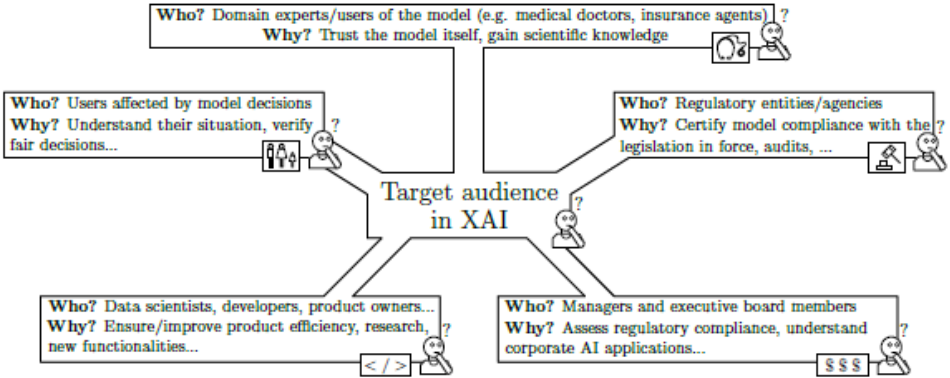
Driven mostly by ethical and legal concerns, **XAI** refers to research into making models and their predictions more understandable by stakeholders such as end-users, domain experts, developers, managers and auditors.

Interpretability

A model is **interpretable** to the extent that humans can gain insight into what has been learned.

Explainability

A prediction is **explainable** to the extent that humans can understand how or why a particular outcome has been predicted by a model.


Who? Domain experts/users of the model (e.g. medical doctors, insurance agents)
Why? Trust the model itself, gain scientific knowledge
Who? Users affected by model decisions
Why? Understand their situation, verify fair decisions...
Who? Regulatory entities/agencies
Why? Certify model compliance with the legislation in force, audits, ...
Target audience in XAI
Who? Data scientists, developers, product owners...
Why? Ensure/improve product efficiency, research, new functionalities...
< / >
Who? Managers and executive board members
Why? Assess regulatory compliance, understand corporate AI applications...
$ $ $

How Interpretable/Explainable?

kNN

- Interpretability does not apply to kNN. Why not?
- kNN is generally thought to be explainable. But when might it not be?

Decision Trees

- DTs are generally thought to be interpretable. But when might they not be?
- DTs are generally thought to be explainable.

Ensembles

- Do you think an ensemble is an interpretable model?
- Do you think that a prediction from an ensemble is easily explainable?

Linear Models

- Linear Regression/Logistic Regression are simple models. People think that the coefficients show feature importance. But this oversimplifies. Why?
- Are their predictions explainable?

Large n

- Large n lessens interpretability/explainability.
- Feature Selection (including l_1 regularization of parametric models) can reduce the number of features and improve XAI.
- Dimensionality Reduction, e.g. PCA, can also be used to reduce the number of features. But it worsens XAI. Why?

Neural Networks are Black Boxes

- In general, neural models are **not interpretable**.
- In general, their predictions are **not explainable**.

Visualizations of Convolutional Neural Networks

- Interpretability:
  - Visualizations of the activations of convolutional and pooling layers.
  - Visualizations of the inputs that convolutional layers are receptive to.
- Explainability:
  - Visualizations of heatmaps that show parts of an image that most contribute to a classification.

Spurious Correlations

- A **spurious correlation** is a correlation that exists between a characteristic of the training examples and the target but where the characteristic is, in reality, not relevant to the task.
- Some spurious correlations are world-induced — just the way the world is.
- E.g. in images of animals, the habitat in the background may be spuriously correlated with the species of the animal.
- Other spurious correlations are sampling-induced — due to selection bias during sampling or introduced by the cleaning or preprocessing of the training set.
- E.g. in images of skin lesions, the presence of a ruler to show lesion size may be spuriously correlated with malignancy.

Shortcuts

- A model learns a **shortcut** when it uses a spurious correlation as the basis for its decision-making.
- It relies on non-relevant characteristics instead of, or as well as, relevant ones.
- It may give correct answers, but for the wrong reasons.
- The shortcut does not generalize well.

Examples of Shortcuts (some may be urban myths)

task	shortcut
animal species classification	snowy background correlated with husky
animal species classification	grassy background correlated with sheep
animal species classification	overlaid photographer info correlated with horse
US/Russian tank classification	sunny weather conditions correlated with US tanks, cloudy with Russian tanks
detection of pneumonia in chest radiographs	type and location of equipment correlated with pneumonia

Why do ML Algorithms Learn Shortcuts? (Hypotheses)

- Imprecise task formulation — the task is object classification but framed as image classification.
- Occam's Razor — it is easier to learn the shortcut than the relevant correlations.
- If the noise or variance in the relevant characteristics is larger than the noise or variance in the spurious features — it means the non-relevant feature are more predictive than the relevant ones.
- Overparameterized models are even more susceptible — they may learn the shortcut and then use their capacity to memorize those training examples that violate the shortcut, rather than learning the relevant characteristics.

An Active Research Area

How to Detect Shortcuts

- Train multiple models with strong regularization, dropout, early stopping. Non-shortcut examples are ones that are consistently correctly classified.
- Perturb training examples in irrelevant ways (e.g. slight cropping): changes in accuracy may be indicative of shortcuts.
- Use heatmaps to show whether the model focuses on irrelevant characteristics.

How to Avoid Shortcuts

- Mask the parts of training examples that give rise to spurious correlations (e.g. mask backgrounds in object classification).
- Use data augmentation to add training examples to counteract ones likely to give rise to spurious correlations (e.g. combine with arbitrary backgrounds).
- Train a low-capacity model — it is likely to rely on shortcuts. Use it as a shortcut detector — examples it gets correct. Then down-weight those examples in the training set of your real model.