Explainable AI and AI Ethics

Guide to AI and Data Courses · Part II: Branches of AI

Why explain? | Explanation methods | Bias and fairness | Accountability | Regulation | Neighboring courses

Sercan KÜLCÜ · Giresun University · Computer Engineering

© Sercan Külcü. All rights reserved. · Interactive HTML version of these slides:
https://sercankulcu.github.io/files/ai_courses/slides/Chapter_15_Explainable_AI_and_Ethics.html

INTRODUCTION

What Question Does It Ask?

> **When can we trust an AI system's decision, and who is responsible for the outcome?**

This course studies how to **explain** an AI system's decisions in a way people can understand, and how to make the system fair, reliable and accountable. The technical question is "why did the model decide this?"; the ethical question is "whom can this decision harm, and who is responsible?"

At a glance

APPROACH human-centered evaluation	**INPUT** model, its data and context of use	**OUTPUT** explanation, audit report, principles
SUCCESS MEASURE interpretability, fairness, accountability	**MATH LOAD** low-medium	**DATA NEEDS** the model's own data

A Short History

2016
ProPublica publishes an analysis arguing that the COMPAS recidivism score is biased against Black defendants.

2016
Ribeiro, Singh and Guestrin introduce LIME: explaining one prediction with a simple local model.

2017
Lundberg and Lee publish SHAP: feature contributions from game-theoretic Shapley values.

2018
The EU General Data Protection Regulation (GDPR) takes effect, granting a right to contest automated decisions.

2018
The Gender Shades study shows commercial face analysis systems err far more often on darker-skinned women.

2024
The EU AI Act is adopted: systems are classified by risk level.

THE COURSE

Core Concepts

Black-box model
A model whose inner workings people cannot follow.

Interpretability
A model readable by design: decision tree, linear.

Local explanation
Explaining one prediction: LIME, SHAP values.

Global explanation
The model's overall behavior: feature importance, PDP.

Counterfactual
"Had your income been this much, the loan would be approved."

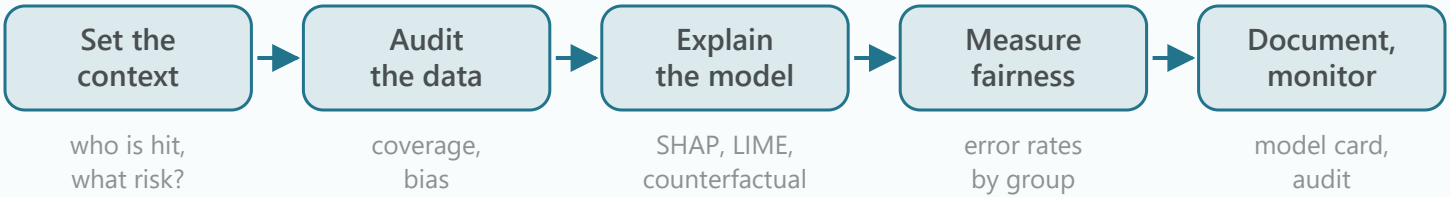
Data bias
Imbalance in training data carried into decisions.

Fairness metrics
Demographic parity, equal opportunity, calibration.

Accountability
Auditing, logging and ownership of a decision.

THE COURSE

A Typical Workflow



Example: A bank's credit model rejects an application. SHAP values show the rejection comes mostly from the debt ratio; a comparison across groups reveals markedly higher rejection rates in some neighborhoods. The team removes the feature that indirectly encodes neighborhood and documents the decision in the model card.

THE COURSE

What Does the Course Cover?

1. Interpretable models: linear models and trees
2. Feature importance and partial dependence (PDP, ICE)
3. Local explanations: LIME
4. Shapley values and SHAP
5. Saliency maps for deep networks
6. Counterfactual explanations and recourse
7. Evaluating explanations: faithfulness and stability
8. Fairness metrics and the conflicts between them
9. Bias mitigation: pre-, in- and post-processing
10. Privacy: anonymization and differential privacy
11. Accountability: model cards, datasheets, audits
12. Regulation: GDPR and the EU AI Act

Skills you will gain

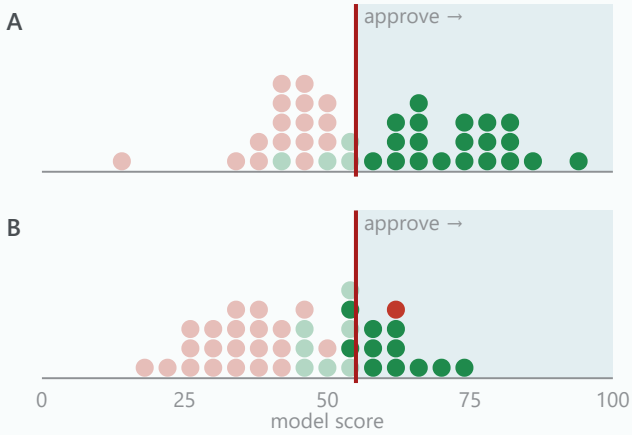
explaining models with SHAP and LIME | measuring group fairness with Fairlearn | generating counterfactual explanations | writing a model card | assessing the risks of a system

A typical semester; the order and coverage vary between universities.

TRY IT

Is the Credit Model Fair to Both Groups?

The model scores each application from 0 to 100; scores above the threshold are approved. Group B scores lower because of historical data. Move the thresholds and watch how two fairness metrics change.



Shared threshold | Per-group threshold

Group A threshold = **55**

Group B threshold = **55**

Equalize approval rates | Equalize opportunity

	approved	repayers approved (TPR)
Group A	50%	83%
Group B	30%	61%

Demographic parity gap: **20 pts**
Equal opportunity gap: **22 pts**
Accuracy: 85%

The same threshold treats the two groups differently; closing one gap opens the other. The course teaches making these measurements, stating openly which fairness definition was chosen, and making the model auditable.

CONTEXT

Prerequisites

Courses worth knowing before this one, and why they are needed.

Course	Why it is needed
Programming	training models in Python and analyzing results
Algorithmic Game Theory	the Shapley value: the theoretical root of SHAP explanations
Philosophy or law	ethical theories, personal data rights and the notion of responsibility

CONTEXT

Are You Ready for This Course?

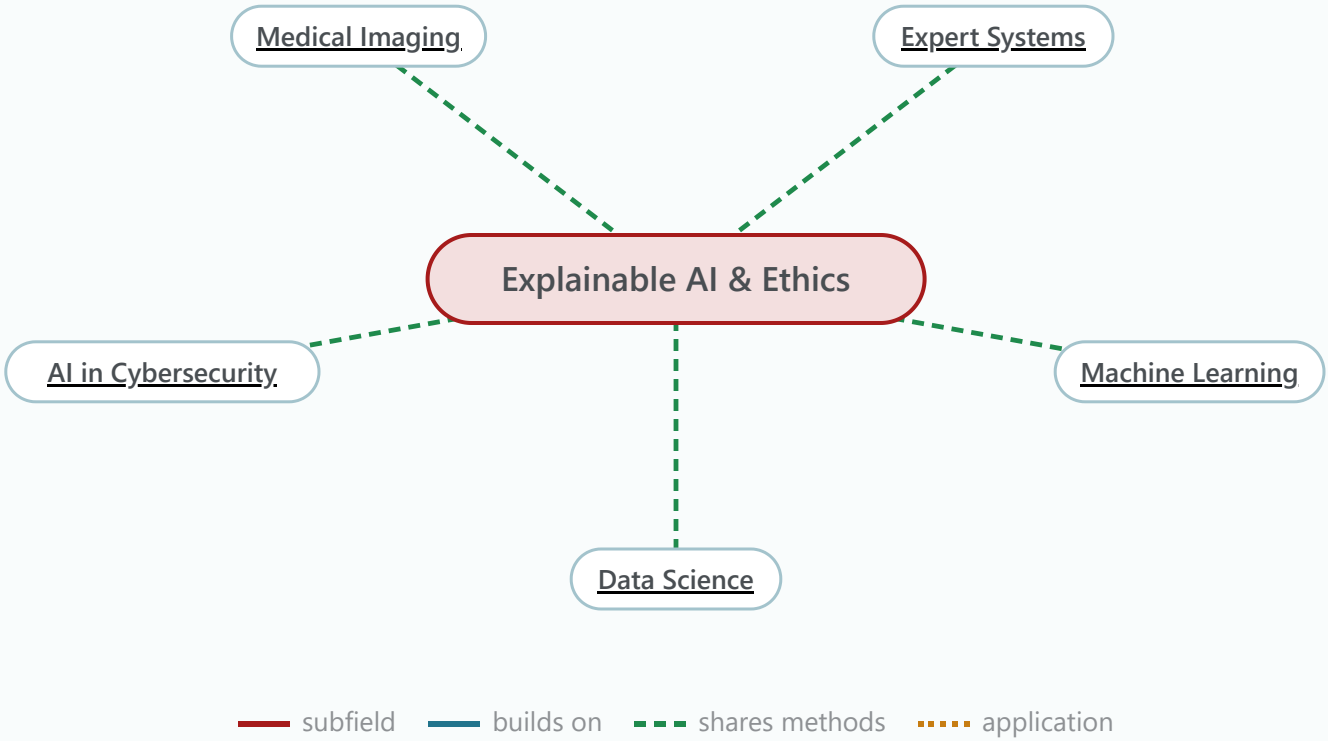
Ask yourself; for every "no", the right-hand column shows where to start.

Ask yourself	If not, start with
Can you train a classifier in Python and evaluate it on a test set?	Machine Learning
Can you compute false positive and false negative rates?	A short reading on confusion matrices and classification metrics
Can you explain conditional probability and base rates?	Probability and Statistics
Can you explain what the Shapley value means in game theory?	The Shapley value topic of the Algorithmic Game Theory course

If you can say "yes" to most of these, you are ready; one or two gaps can be closed during the semester.

CONTEXT

Neighbouring Courses



OVERLAP

Overlapping Courses

Relation	Course	Shared ground
shares methods	Shares methods with Machine Learning	model explanation, feature importance
shares methods	Shares methods with Expert Systems and Knowledge Representation	explainable decisions through readable rules
shares methods	Shares methods with Medical Image Processing	explainability and accountability in diagnosis
shares methods	Shares methods with Data Science	data bias, fairness metrics
shares methods	Shares methods with AI in Cybersecurity	robustness, adversarial examples

DIFFERENCES

Where It Differs (1/2)

Criterion	Explainable AI & Ethics	Machine Learning	Data Science
Core question	When can we trust an AI system's decision, and who is responsible for the outcome?	How does a program generalize from examples?	How do we ask the right questions of data and turn the answers into decisions?
Approach	human-centered evaluation	statistical learning	question, data, analysis, story
Input	model, its data and context of use	labeled or unlabeled examples	raw, messy data
Output	explanation, audit report, principles	a model that makes predictions	report, visualization, model
Success measure	interpretability, fairness, accountability	error on unseen data	impact on decisions, accuracy
Math load	low-medium	high: linear algebra, probability, optimization	medium: statistics
Data needs	the model's own data	moderate to large	moderate to large

- **Machine Learning:** ML focuses on the model predicting correctly; this course asks why it predicts and whom it affects, and how.
- **Data Science:** Data science draws conclusions from data; this course checks whether those conclusions are biased, fair or explainable.

DIFFERENCES

Where It Differs (2/2)

Criterion	Explainable AI & Ethics	AI in Cybersecurity	Expert Systems
Core question	When can we trust an AI system's decision, and who is responsible for the outcome?	How does AI catch attacks, and how is it attacked itself?	How is an expert's knowledge written down and used for inference?
Approach	human-centered evaluation	security knowledge plus ML	symbolic logic and rules
Input	model, its data and context of use	network traffic, logs, files	facts and expert rules
Output	explanation, audit report, principles	alert, class, risk score	conclusion with justification
Success measure	interpretability, fairness, accountability	detection rate, false alarm rate	expert agreement, explainability
Math load	low-medium	medium	low: propositional and predicate logic
Data needs	the model's own data	imbalanced and constantly changing	little; a knowledge base

- **AI in Cybersecurity:** Cybersecurity protects against attackers; this course studies the harm a system does to people even without an attack.
- **Expert Systems:** Expert systems are already transparent through their rules; this course seeks ways to explain black-box models after the fact.

PRACTICE

Applications

Credit decisions
Explaining a rejection to the applicant.

Medical diagnosis
Showing the physician what the model looked at.

Hiring
Gender and age bias in résumé screening.

Justice and public sector
Auditing whether risk scores are fair.

Face recognition
Measuring error gaps between groups.

Regulatory compliance
Documentation for the EU AI Act and KVKK.

PRACTICE

Risks and Ethics

Misleading explanations

LIME and SHAP are approximations; an adversarial model can look innocent to the explainer while hiding its real decision rule.

Ethics washing

An ethics principles document or a single fairness score can serve as a showcase in place of real auditing.

The wrong fairness metric

Choosing a metric is a value judgement: COMPAS could look fair under one metric and biased under another.

Over-trusting explanations

A plausible-looking explanation can lead users to follow a wrong decision without question.

PRACTICE

Tools and Datasets

Tools and libraries

SHAP | LIME | Captum | InterpretML | Fairlearn | AI Fairness 360 | What-If Tool

Well-known datasets

COMPAS (ProPublica) | Adult (income) | German Credit | CelebA | Pilot Parliaments Benchmark

In this course datasets are often used **to expose a problem**: COMPAS and Adult are the classic examples of how bias is measured.

PRACTICE

Common Misconceptions

Common belief	In fact
Removing the protected attribute (gender, race) is enough.	Proxy features such as postal code or name can carry the same information; the bias remains.
An algorithm is neutral because it works with numbers.	A model learns the bias in past decisions from the data and scales it up.
An explanation shows how the model really works.	Methods such as LIME and SHAP are approximations; the explanation itself can mislead.
All fairness metrics can be satisfied at once.	When base rates differ between groups, calibration and equal error rates usually cannot hold together.

NEXT STEPS

After This Course

Courses that build on or extend this one

This course is mostly an end point: it builds on the others.

Related careers

Responsible AI specialist | Model auditor | AI policy and compliance specialist | Data scientist | Researcher

This course adds a lens to every applied course: explaining a diagnosis in medical image processing, fooling a model in cybersecurity and monitoring a model in MLOps rest on the same questions. On the technical side you can continue with causal inference and adversarial robustness; on the social side, with law and policy.

NEXT STEPS

Further Reading

- Molnar, *Interpretable Machine Learning*, 2nd ed., self-published (online), 2022.
- Barocas, Hardt and Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*, MIT Press, 2023.
- O'Neil, *Weapons of Math Destruction*, Crown, 2016.
- Ribeiro, Singh and Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier", *KDD*, 2016.
- Lundberg and Lee, "A Unified Approach to Interpreting Model Predictions", *NeurIPS*, 2017.

The End

Chapter 15: Explainable AI and AI Ethics

← Previous chapter · Map of the field · Next chapter →