Introduction to the course

2024-2025

Explainable and Trustworthy AI

Eliana Pastor
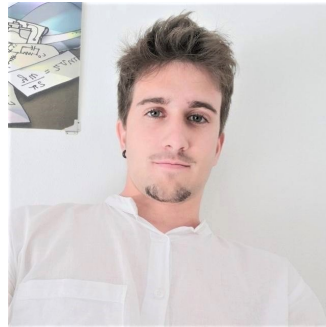
Teaching staff

Eliana Pastor

Gabriele Ciravegna

Eleonora Poeta

Course topics in a glance

- Trustworthy AI: Notions and Requirements
- Taxonomy of Explainable AI (XAI)
- Explainable by design models
- Pre-modeling explainability
- In-modeling explainability
- Post-hoc model agnostic approaches
- Concept-based explainable AI
- Evaluation of explanations
- Attention-based explainability
- Adversarial Attacks
- Counterfactual explanations

Material

Announcements on the official «teaching portal» private page https://didattica.polito.it/ or through institutional email

Explainable and trustworthy AI

01HFOOV, 01HFOSM

A.A. 2024/25

Slides, texts of the practices, and all other materials are available on the public page of the course:

https://dbdmg.polito.it/dbdmg_web/2025/explainable-and-trustworthy-ai-2024-2025/

📁 Courses

Explainable and trustworthy AI (2024/2025)

General Information

SSD: ING-INF/05

CFU: 6

Lecturer: Eliana Pastor

Teaching Staff: Gabriele Ciravegna, Eleonora Poeta

Course structure

The course includes

- Lectures

- Hands-on and exercises

- Laboratories
    - Experimental activities
    - Practical analysis and evaluations of the methods covered in the course

- Lab practises will start from the third week

Schedule

- **Monday**

  8:30-10:00

  classroom 11T

- **Thursday**

  8:30-10:00

  classroom 5T

- **Friday**

  8:30-10:00

  classroom 2I

No fixed slot distinction between Lecture and Lab Practise

Exam

- Written exam
    - Main definitions and concepts of Explainable and Trustworthy AI
    - Explanation techniques and their main characteristics
    - Main libraries implementing explanation methods

- Group project (2-3 students)
    - Implement and evaluate a complete data science pipeline and its explanation
    - Design and evaluate explanation methods
    - Presentation in written form

The detailed exam rules will be soon published in the web page of the course

Project goal and required objectives

- **Literature review**: systematic review of the works related to the project topics

- **Research gap**: identification of research gaps of the problem given the current literature

- **Methodology and Implementation**: propose and implement a possible solution addressing such research gaps

- **Analysis**: assess the proposed solution and analyze and discuss the outcomes