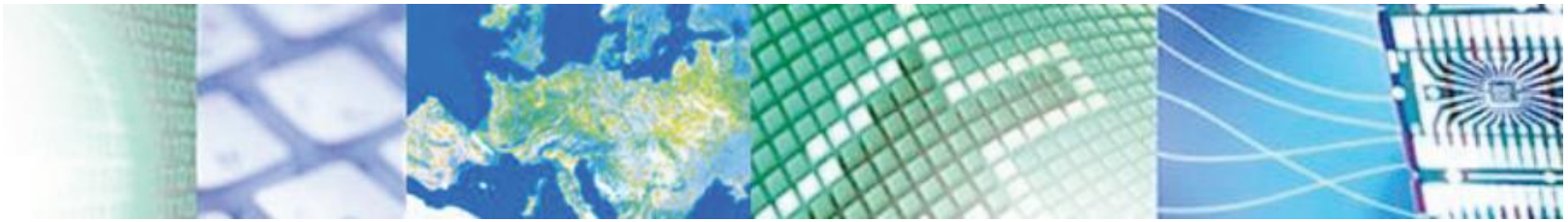


CHIST-ERA Projects Seminar 2023
Explainable Machine Learning-based Artificial Intelligence (XAI)

April 04, 2023

Programme co-funded by the
EUROPEAN UNION

Introduction: Definition of the Topic

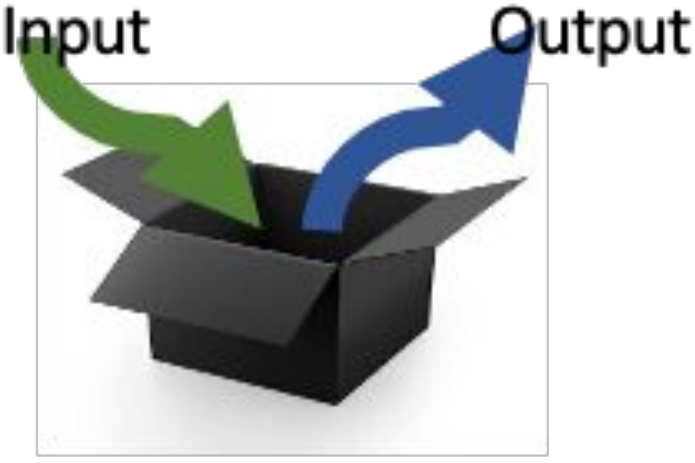
**Most AI methods are "black boxes", i.e.,
it's difficult to know how and why they work.**

**XAI are tools to explain the models and their decision,
so that humans can trust them.**

Modern AI models, such as deep learning, are trained in an end-to-end manner from "big data" → it's unclear what exactly the model is learning.

Important keywords:

- Prediction / Classification
- Data-driven
- Trust
- Transparency
- Interpretability of the models
- Human-centric
- Personalization
- Explanation negotiation/argumentation
- Neuro-Symbolic integration
- Ethics and legal for xAI



Introduction: Projects of the Topic

ANTIDOTE: Provide a unified computational framework for jointly learning clinical predictions and the associated argumentative justifications.
CausalXRL: Learn two-level Causal model of environment to enable Reinforcement Learning agent to suggest eXplainable actions in critical environments.
CIMPLE: Automatically detect misinformation & generate creative explanation response
COHERENT: Generate and combine explanations and decide when to deliver them during robotic collaboration tasks.
EXPECTATION: Generate personalized explanations based on heterogeneous knowledge during iterative interactions with users and agents.
GraphNEx: Graph neural networks for XAI & user-guided explainability
INFORM: Interpretability of DNNs in oncology applications on data extracted from medical images (radiomics), and evaluation of interpretability methods
iSEE: Retrieve, personalise and reuse explanation strategies on different use case based on past explanation experiences evaluation against persona intents
MUCCA: Quantifying strengths and solving weaknesses of new and state of the art XAI methods using heterogeneous use cases
SAI: Decentralised ML-based AI in human social complex systems
XAIFace: Explain black box facial recognition methods and fight against biased results.
XPM: How explanations support decision making in predictive maintenance systems

Major Achievements and Outputs

XAI algorithms & methods

- Resource-efficient training
- Incorporated ontologies to standardize and shape explanations
- Approaches for generating or co-creating personalized explanations
- Improved reusability of existing XAI algorithms

Platforms & Frameworks & Toolkits

- Conversational dialogues with personalising explanations
- Symbolic knowledge extraction & injection
- Simulators for large-scale decentralised AI evaluation
- Explanation strategies design, recommendation and evaluation platform

Evaluating XAI methods

- Technological evaluation of the reliability of explanations
- Subjective evaluation of the whole experiences of explanations
- User-guided explainability
- User-centric metrics for explanations

Upcoming Challenges and Needs

Make the user central in the XAI process

Explanation personalization according to users' profiles & needs

Subjectivity in the explanation evaluation

Generalization of XAI methods to different domains

Acceptance and implementation of AI systems

Challenging for end-users to trust developed AI models

XAI standardisation guiding organisations using AI systems

Compliance to emerging AI regulation and guidelines (ALTAI)

Explanation(s) alignment (in case of disagreement) in distributed envs

Ontologies (Knowledge representation) alignment across different parties/actors

Integration of different nature/data type explanations

Possible Roadmap

Multidisciplinary approach, incorporating different perspectives and techniques inspired by existing state-of-the-art work, and sharing the knowledge and experience gained through the project with other researchers.

Experiments with users to iteratively get their feedback, for example, design an interactive decision support system to evaluate different types of explanations for different types of users.
Investigating different levels of explanations based on user expertise.

Exploiting connection between the predictions and the underlying dynamics governing the system, which is generally assumed to be well-understood by the expert and can therefore provide common ground between the AI and the human.

Understand explainability at the collective level in decentralised, collaborative AI.

Define a legal and ethical framework for XAI based systems.

Expected impact:

overall higher acceptance of AI systems in all areas of human life, and more knowledge created through use of AI

Role of the CHIST-ERA Support

Very interesting and relevant topic

(plus community-driven decision process for defining them)

Bringing the researchers together

Facilitate the contact between researchers of different projects to identified complementaries (ex: new use case for XAI algorithm)

One of the few programmes looking for long-term research ideas

Idea: propose rules concerning different funding categories (e.g., min travel budget), and properly communicate them to the funding agencies

Facilitating follow-ups

Complementarity with Horizon Europe

Focus on small medium organizations and consortiums with a small number of partners

Topics side (XAI):

- Focus on real world applications (including metrics applicable for certification)
- Add causality as a key topic to research
- Bring together XAI and Human Computer Interaction

Responsible Research & Innovation

Integrated good practices for RRI

- Publish all papers in institutional repositories (open access)
- Share the data and software in zenodo/other repositories
- Conducting human experiments under the supervision of ethical committees
- Include ethics in the dimensions of XAI

Major hurdles to implement RRI

- Gender balance in technical domains is difficult due to the low percentage of women in computer science, robotics and AI
- Predatory journals are not well identified and create an unfair environment to develop science
- Cost of open-access journals → we can publish the author version in your institutional repository and publish without paying

Open Science

Public dissemination through scientific publications and conferences

- Free online access or open gold model

New data collected is released as public accessible databases

Open source and reproducible research platforms are used

Data: FAIR principles are followed
(Findable, Accessible, Interoperable, Reusable)

Explore EU platforms: OpenAIRE, Open Research Europe

Technology Transfer

Establishing **community** and social network of XAI researchers

Open Source **platform** and European XAI **compliance certification** framework

Cross-application assessment of available approaches to XAI

Find a **common representation** for heterogeneous explanation methods

Integration of novel solutions into existing commercial solutions

Actions should involve all CHIST-ERA member countries:

- Identify transferable technologies
- Survey of potentially interested stakeholders
- Participate/organize events (workshops / training schools/ challenges)
- Send invitations to others project members

Questions

Questions ?